Desenvolvimento de pipeline de bioinformática para validação *in silico* de marcadores microssatélites utilizando dados de genotipagem em larga escala (ddRAD-seq) em plantas de *Agave*

Palavras-Chave: sisal, SSR, marcadores moleculares.

João Vitor Rodrigues Mio[1], Marina Püpke Marone[2] (coorientadora), Marcelo Falsarella Carazzolle[1] (coorientador), Gonçalo Amarante Guimarães Pereira[1] (orientador)

[1]Instituto de Biologia, Unicamp.

[2]Leibniz Institute of Plant Genetics and Crop Plant Research – Gatersleben

1. INTRODUÇÃO.

Agaves são plantas resistentes à seca que desempenham um papel importante na produção de fibras de sisal e possuem um potencial significativo na produção de açúcar solúvel e biomassa para bioenergia. Os agaves são uma alternativa para regiões semiáridas como o sertão Nordestino, 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 para obter mais indivíduos para o banco de germoplasma e genotipá-los. 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 dois genomas parciais de *Agave*, obtivemos sequências de SSR e selecionaremos sequências polimórficas baseado nos dados de ddRAD-seq de 95 indivíduos. Testes em bancada para validação de primers que amplificam SSR são laboriosos e demandam muito tempo. Com este pipeline, será

possível realizar uma etapa preliminar de validação in sílico, o que permitirá economizar recursos.

2. METODOLOGIA.

2.1 Dados utilizados.

Os dados para identificação dos locos SSRs são genomas draft de *A. sisalana* e híbrido 11648 (H11648), que possuem regiões gênicas, incluindo promotor e íntrons. Os dados de genotipagem foram obtidos através da técnica ddRAD-seq. Um total de 95 indivíduos foi sequenciado, sendo 77 *A. sisalana* e 18 híbrido 11648.

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.

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). O software Samtools (DANECEK *et al.*, 2021) foi utilizado para converter os arquivos SAM e o software BEDTools (QUINLAN; HALL, 2010) foi utilizado para conversão em formato BED. Para selecionar somente os reads de cada indivíduo que têm sobreposição com as regiões de SSR, foi utilizado o módulo subtract presente no BEDTools.

A partir dessas regiões, foi desenvolvido um código em Python que identifica os reads alinhados que atravessam regiões de SSRs (Figura 1). Este código gera arquivos com coordenadas para início e fim das regiões que flanqueiam a sobreposição do read com o SSR. Em seguida, foi desenvolvido outro script em Python (ainda em processo de otimização) para identificar e contar as regiões polimórficas entre os indivíduos. O código permite que sejam escolhidos e ajustados os critérios para definir quais reads são considerados polimórficos dentro das regiões de SSR. Com base nessas informações, serão selecionados os primers com parâmetros adequados para os futuros testes em bancada.

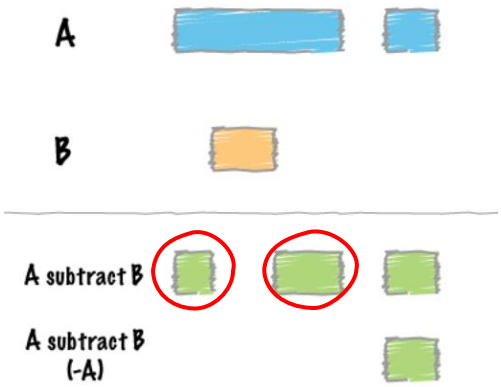

Figura 1. Representação esquemática do módulo subtract utilizado como abordagem na busca de reads ddRAD contemplados com as regiões de SSR. "A" representa os reads de ddRAD-seq e "B" as regiões SSRs. Os círculos vermelhos representam início e fim das regiões que flanqueiam a sobreposição do read com o SSR.
Disponível em: <https://bedtools.readthedocs.io/en/latest/content/tools/subtract.html>.

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

A obtenção dos locos SSRs foi realizada a partir de montagens das regiões gênicas de *A. sisalana* e do híbrido 11648. No entanto, apenas uma parte das sequências de ddRAD que se alinharam a esses locos está sendo considerada. 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.

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**
Agave sisalana	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, realizamos a análise para identificar os reads de ddRAD-

seq que atravessam regiões de SSR. Ao final do processo, foram obtidos 24.271 arquivos para H11648 e 38.782 para *A. sisalana*.

Embora existam milhares de regiões identificadas na etapa anterior, ainda é necessário identificar se os reads que atravessam um determinado SSR são polimórficos entre indivíduos. Para isso, o segundo código desenvolvido retorna somente aqueles arquivos de saída que possuem inserções ou deleções no alinhamento dos reads ao SSR. Foi definido um mínimo de dois reads em cada indivíduo para considerar uma região como polimórfica. Após executar essa filtragem, obtivemos um total de 623 arquivos de saída para *A. sisalana* e 714 para o H11648. Os próximos passos são converter para formato tabular, tornando as análises subsequentes mais eficientes, e selecionar os primers que serão testados em bancada.

Uma outra proposta do projeto é 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). Em seguida, identificaremos os locos que são exclusivos e polimórficos da estratégia *de novo*.

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**
A. sisalana	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 714 regiões polimórficas em H11648 e 623 em *A. sisalana*. Além disso, está em andamento uma análise *de novo*, a partir da qual vamos buscar por regiões SSR que não foram identificadas com a primeira estratégia. A partir da lista final de locos polimórficos, selecionaremos os primers correspondentes para serem testadas em bancada utilizando os mesmos 95 indivíduos. O pipeline desenvolvido permite aceleração da validação de primers que amplificam locos SSRs e redução de custos para os testes em bancada. Mais

especificamente, os resultados vão gerar primers que serão utilizados para distinguir indivíduos de *A. sisalana* e H11648 e impulsionar 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**, v. 27, n. 3, p. 513–520, ago. 1997.

CATCHEN, J.; HOHENLOHE, P. A.; BASSHAM, S.; AMORES, A.; CRESKO, W. A. Stacks: An Analysis Tool Set for Population Genomics. **Molecular Ecology**, v. 22, n. 11, p. 3124–3140, jun. 2013.

DANECEK, P.; BONFIELD, J. K.; LIDDLE, J.; MARSHALL, J.; OHAN, V.; POLLARD, M. O.; WHITWHAM, A.; KEANE, T.; MCCARTHY, S. A.; DAVIES, R. M.; LI, H. Twelve Years of SAMtools and BCFtools. **GigaScience**, v. 10, n. 2, p. giab008, 29 jan. 2021.

IDREES, M.; IRSHAD, M. Molecular Markers in Plants for Analysis of Genetic Diversity: A Review. n. 1, p. 28, 2014.

LI, H.; DURBIN, R. Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform. **Bioinformatics**, v. 25, n. 14, p. 1754–1760, 15 jul. 2009.

QUINLAN, A. R.; HALL, I. M. BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. **Bioinformatics**, v. 26, n. 6, p. 841–842, 15 mar. 2010.

SENAN, S.; KIZHAKAYIL, D.; SASIKUMAR, B.; SHEEJA, T. E. Methods for Development of Microsatellite Markers: An Overview. **Notulae Scientia Biologicae**, v. 6, n. 1, p. 1–13, 12 mar. 2014.

THIEL, T.; MICHALEK, W.; VARSHNEY, R.; GRANER, A. Exploiting EST Databases for the Development and Characterization of Gene-Derived SSR-Markers in Barley (Hordeum Vulgare L.). **Theoretical and Applied Genetics**, v. 106, n. 3, p. 411–422, fev. 2003.

UNTERGASSER, A.; CUTCUTACHE, I.; KORESSAAR, T.; YE, J.; FAIRCLOTH, B. C.; REMM, M.; ROZEN, S. G. Primer3—New Capabilities and Interfaces. **Nucleic Acids Research**, v. 40, n. 15, p. e115–e115, ago. 2012.