

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE CIÊNCIAS AGRÁRIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRICULTURA E INFORMAÇÕES GEOESPACIAIS

MÁRIO LÚCIO MARTINS ROCHA

Caracterização *in silico* de STRs em genomas de *Aspergillus niger* visando ao desenvolvimento de marcadores moleculares para cepas de bioinsumos agrícolas

MÁRIO LÚCIO MARTINS ROCHA

Caracterização *in silico* de STRs em genomas de *Aspergillus niger* visando ao desenvolvimento de marcadores moleculares para cepas de bioinsumos agrícolas

Dissertação apresentada ao Programa de Pós-graduação em Agricultura e Informações Geoespaciais da Universidade Federal de Uberlândia como requisito para obtenção do título de mestre.

Área de concentração: Informações geoespaciais e tecnologias aplicadas à produção agrícola

Orientador: Gilberto de Oliveira Mendes

Coorientadores: Ana Carolina Silva Siquieroli

Murillo Guimarães Carneiro

Monte Carmelo

2026

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

R672 2026	Rocha, Mário Lúcio Martins, 1977- Caracterização in silico de STRs em genomas de <i>Aspergillus niger</i> visando ao desenvolvimento de marcadores moleculares para cepas de bioinsumos agrícolas [recurso eletrônico] / Mário Lúcio Martins Rocha. - 2026. Orientador: Gilberto de Oliveira Mendes. Coorientadora: Ana Carolina Silva Siquieroli. Coorientador: Murillo Guimarães Carneiro. Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Agricultura e Informações Geoespaciais. Modo de acesso: Internet. DOI http://doi.org/10.14393/ufu.di.2026.552 Inclui bibliografia. Inclui ilustrações. 1. Agronomia. I. Mendes, Gilberto de Oliveira, 1983-, (Orient.). II. Siquieroli, Ana Carolina Silva, 1982-, (Coorient.). III. Carneiro, Murillo Guimarães, 1988-, (Coorient.). IV. Universidade Federal de Uberlândia. Pós-graduação em Agricultura e Informações Geoespaciais. V. Título. CDU: 631
--------------	---

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
 Coordenação do Programa de Pós-Graduação em Agricultura e Informações
 Geoespaciais
 Rodovia LMG 746, Km 01, s/nº, Bloco 1AMC, Sala 1A202, Monte Carmelo-MG, CEP 38.500-000
 Telefone: (34) 3810-1033 - ppgaig@iciag.ufu.br



ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Agricultura e Informações Geoespaciais				
Defesa de:	Dissertação de Mestrado Acadêmico				
Data:	30/04/2026	Hora de início:	14:00	Hora de encerramento:	16:15
Matrícula do Discente:	32322AIG011				
Nome do Discente:	Mário Lúcio Martins Rocha				
Título do Trabalho:	Caracterização <i>in silico</i> de STRs em genomas de <i>Aspergillus niger</i> visando ao desenvolvimento de marcadores moleculares para cepas de bioinsumos agrícolas				
Área de concentração:	Informações geoespaciais e tecnologias aplicadas à produção agrícola				
Linha de pesquisa:	Sistemas integrados de produção vegeta				
Projeto de Pesquisa de vinculação:	Processos e produtos microbianos aplicados à produção vegetal				

Reuniu-se em sala virtual a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Agricultura e Informações Geoespaciais, composta pelo Dr. Gilberto de Oliveira Mendes (Universidade Federal de Uberlândia-UFU), orientador do candidato, Dr. André Luiz Firmino (Universidade Federal de Uberlândia-UFU) e Dr. Tomás Gomes Reis Veloso (Universidade Federal de Viçosa -UFV).

Iniciando os trabalhos o presidente da mesa, Dr. Gilberto de Oliveira Mendes, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao discente a palavra para a exposição do seu trabalho. A duração da apresentação do discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado.

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.



Documento assinado eletronicamente por **Gilberto de Oliveira Mendes, Professor(a) do Magistério Superior**, em 30/04/2026, às 16:19, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **André Luiz Firmino, Professor(a) do Magistério Superior**, em 14/05/2026, às 12:18, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **TOMAS GOMES REIS VELOSO, Usuário Externo**, em 15/07/2026, às 17:15, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **7269971** e o código CRC **B0717377**.

RESUMO

A microbiota do solo, especialmente os fungos, desempenham papel central na promoção do crescimento vegetal e na ciclagem de nutrientes. Entre eles, *Aspergillus niger* destaca-se por seu potencial biotecnológico, contribuindo para a solubilização de fosfatos e o estímulo ao desenvolvimento das plantas, sendo amplamente utilizado como bioinsumo. No entanto, a aplicação desses microrganismos requer o desenvolvimento de estratégias que possibilitem o rastreamento preciso e confiável do destino e do comportamento das cepas introduzidas no ambiente. Nesse contexto, o presente estudo visa aprofundar a caracterização genômica de *A. niger*, com foco em microssatélites (short tandem repeats, STRs), para a identificação de marcadores genéticos com alto poder discriminatório entre cepas, por meio de um fluxo integrado de detecção, avaliação de variabilidade e validação *in silico*. Além de ampliar a caracterização desses marcadores, a estratégia proposta busca oferecer uma abordagem computacional eficiente, reproduzível, escalável e de baixo custo, reduzindo a dependência de etapas experimentais nas fases preliminares de desenvolvimento dos marcadores moleculares.

Palavras-chave: bioinformática, biologia molecular, fungo, microssatélites, SSR, sustentabilidade.

ABSTRACT

Soil microbiota, particularly fungi, play a central role in plant growth promotion and nutrient cycling. Among them, *Aspergillus niger* stands out due to its biotechnological potential, contributing to phosphate solubilization and enhancing plant development. However, the application of such microorganisms requires robust strategies to enable precise and reliable tracking of the fate and behavior of introduced strains under environmental conditions. In this context, this study advances the genomic characterization of *A. niger* by focusing on microsatellites (short tandem repeats, STRs) for the identification of genetic markers with high discriminatory power among strains, through an integrated in silico workflow of detection, variability assessment and validation. In addition to expanding the characterization of these markers, the proposed strategy provides an efficient, reproducible, scalable, and cost-effective computational framework that reduces reliance on experimental procedures during the early stages of molecular marker development.

Keywords: bioinformatics, molecular biology, fungus, microsatellites, SSR, sustainability.

SUMÁRIO

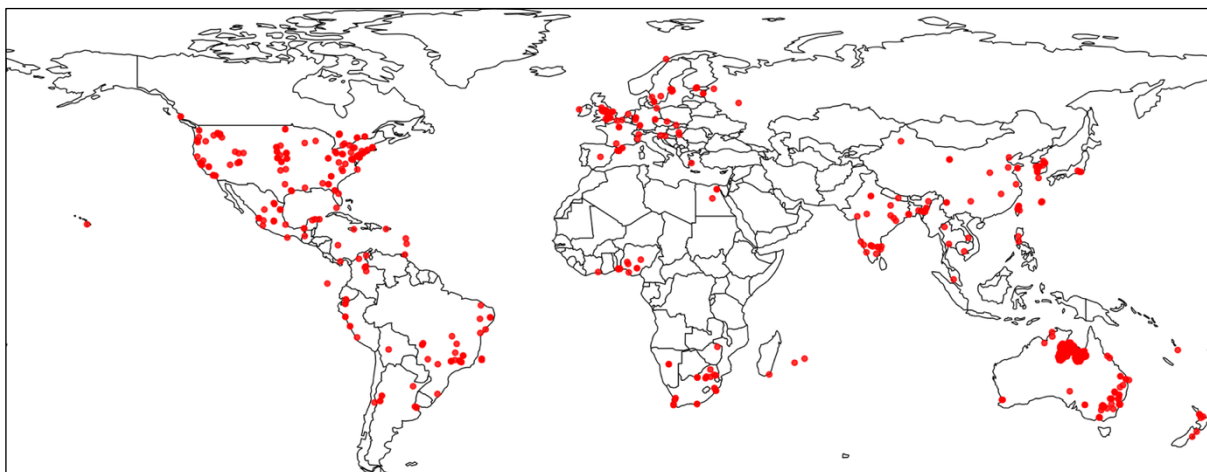
INTRODUÇÃO	9
MATERIAL E MÉTODOS	10
RESULTADOS E DISCUSSÃO	13
CONCLUSÕES.....	21
REFERÊNCIAS	23
APÊNDICES	24

INTRODUÇÃO

O uso consciente do solo é fundamental para o desenvolvimento de uma agricultura sustentável e resiliente. O solo não é apenas um suporte físico, mas um ambiente vivo, dinâmico, rico em interações biológicas essenciais ao crescimento vegetal. A microbiota do solo desempenha papel central na ciclagem de nutrientes, na solubilização de minerais, na promoção do crescimento radicular e na proteção contra patógenos [1]. Valorizar e potencializar esses mecanismos naturais é uma estratégia indispensável frente à necessidade de reduzir a dependência de insumos sintéticos e preservar a saúde dos agroecossistemas.

A engenharia do microbioma da rizosfera é considerada uma ferramenta estratégica para fortalecer a saúde das plantas e aumentar a sustentabilidade agrícola [2]. Nesse cenário, os bioinsumos microbianos, especialmente os fúngicos, ganham destaque como alternativas eficazes e ambientalmente seguras. Os fungos solubilizadores de fósforo representam uma alternativa sustentável para aumentar a disponibilidade de fósforo no solo, melhorando a nutrição e produtividade das plantas e reduzir os impactos ambientais da agricultura convencional [3]. Em particular, o fungo *Aspergillus niger* apresenta um potencial biotecnológico relevante. A aplicação de *A. niger* pode aumentar a disponibilidade de nutrientes e estimular o desenvolvimento inicial das plantas, favorecendo a produção agrícola de forma sustentável [4]. A distribuição espacial global do *A. niger* evidencia que este fungo possui ampla capacidade de adaptação, estando presente em praticamente todas as regiões do planeta, o que potencializa seu papel como ferramenta sustentável para a agricultura conforme mostrado na Figura 1.

Figura 1. Distribuição geográfica do fungo *A. niger*. Mapa ilustrando a ocorrência do organismo em diferentes continentes, evidenciando sua ampla dispersão global.



Fonte dos dados: <https://www.gbif.org/>. Figura: O autor.

A crescente utilização de microrganismos como bioinsumos na agricultura levanta uma questão central ainda pouco resolvida: como rastrear, de forma precisa e rápida, o destino e o comportamento das cepas introduzidas no ambiente. Após a aplicação, raramente se sabe se a cepa inoculada efetivamente se estabeleceu, se sua população aumentou ou diminuiu ao longo do tempo, ou ainda se desempenhou o papel esperado em condições de campo. As avaliações convencionais, baseadas em atributos fenológicos visuais das plantas e em testes agronômicos tradicionais, embora essenciais, são indiretas e não permitem distinguir com segurança o efeito específico da cepa aplicada frente à microbiota nativa. Essa limitação evidencia uma lacuna importante: a ausência de ferramentas moleculares robustas que possibilitem o monitoramento específico e confiável dessas cepas no ambiente, uma vez que o rastreamento de inoculantes no solo permanece um desafio devido à elevada complexidade da microbiota nativa e às limitações dos métodos de monitoramento atualmente empregados [5]. Nesse sentido, este trabalho teve como objetivo desenvolver um pipeline computacional para a identificação e o mapeamento de microsatélites (STRs) em genomas de *A. niger*, visando ao desenvolvimento de marcadores genéticos que subsidiem a discriminação e a rastreabilidade de cepas desse fungo.

MATERIAL E MÉTODOS

A. Origem dos genomas analisados

Este estudo foi conduzido por meio de análises *in silico* de genomas de *A. niger* disponibilizados publicamente pelo National Center for Biotechnology Information (NCBI), detalhados na Tabela 1. Os genomas de *A. niger* tem um tamanho médio de 37,2 milhões de pares de bases.

Tabela 1. Genomas Analisados

Nome da Montagem	Cepa (strain)	Nível de montagem	Tamanho (Mb)
ASM285v2 *	CBS 513.88	Scaffold	34
ASM1928827v1	CBS 554.65	Chromosome	40
ASM2978390v1	KJC3	Chromosome	40
ASM2978392v1	KYF3	Chromosome	37
ASM4765177v1	CCTCC 206047	Complete	35

Genomas de *A. niger* analisados neste estudo. Fonte:

<https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=5061>

* Genoma de referência do táxon.

B. Análise de STRs

A análise de microssatélites (short tandem repeats, STRs) pode ser utilizada de forma complementar ao sequenciamento completo do genoma, otimizando a identificação e caracterização de fungos [6]. Os STRs consistem em motivos curtos que podem variar de 1 a 9 pares de bases (Figura 2). A variabilidade de comprimento e a ocorrência ubíqua de STRs levaram a que fossem amplamente utilizados como marcadores em mapeamento genético [7], [8].

Figura 2. Representação esquemática de STRs em sequências de DNA.

A – Motivo repetitivo (ACTGT) com 5 cópias em tandem.

B – STR com unidade de repetição de 3 pb e 7 cópias.

C – Mesma sequência analisada com unidade de 6 pb, evidenciando sobreposição de padrões, sendo considerado o motivo de menor tamanho.

A	...ACTAGATACGC ACTGT ¹ ACTGT ² ACTGT ³ ACTGT ⁴ TAGTCGTAGACTAGCTACGTAGTCGTAG...
B	...GTACTAGATAGCT GCA ¹ GCA ² GCA ³ GCA ⁴ GCA ⁵ GCA ⁶ GCA ⁷ TGAGTCGATCGCATGCGATGAC...
C	...GTACTAGATAGCT GCAGCA ¹ GCAGCA ² GCAGCA ³ GCATGAGTCGATCGCATGCGATGAC...

Fonte: O autor.

C. Triagem dos STRs

As sequências genômicas foram submetidas à triagem por meio de um algoritmo próprio visando a identificação de repetições perfeitas com unidades variando de mono- a nonanucleotídeos (1 a 9 pares de bases). O algoritmo realiza uma leitura sequencial das bases de nucleotídeos e compara cada base com as próximas, anotando os motivos repetitivos contínuos e registrando a posição, o tipo de motivo correspondente à quantidade de repetições (mono-, di-, tri-, tetra-, penta-, hexa-, hepta-, octo-, e nono nucleotídeos), o número de repetições e o comprimento total do STR. Cada etapa do algoritmo grava as saídas em arquivos do tipo .csv com os valores delimitados por ponto e vírgula (;) o que permite que a próxima fase leia as informações a partir de um ponto de parada ou continuação, bem como o compartilhamento desses dados entre as fases do processo. A primeira parte do algoritmo gera nove arquivos para cada amostra de genoma. Em seguida, é realizado um processo para contar a quantidade de ocorrências desses motivos de mesmo tamanho ao longo de cada amostra de genoma. Os STRs identificados e catalogados em cada genoma foram então comparados entre si para identificar os STRs coincidentes (compartilhados entre os genomas) e STRs exclusivos (presentes em apenas um dos genomas).

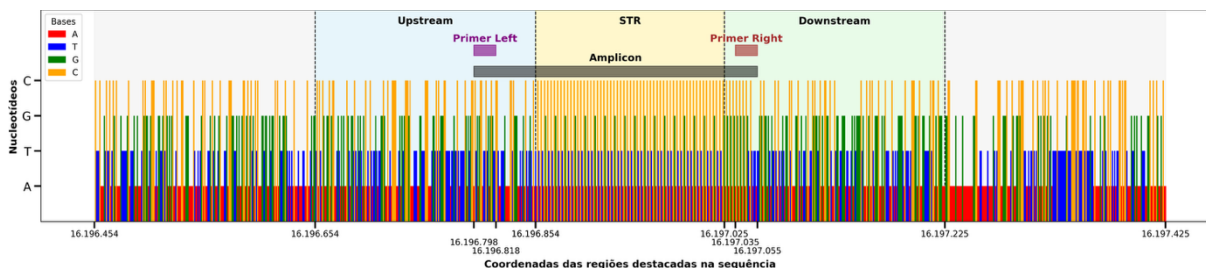
D. Localização das regiões flangeadoras dos STRs

Após a identificação dos STRs exclusivos nos genomas analisados, foram determinadas as regiões flangeadoras associadas a cada STR detectado. Essas regiões correspondem a trechos de DNA posicionados logo antes (5') e depois (3') do motivo repetitivo. Sua função é servir como pontos de ancoragem para os primers, garantindo que a amplificação dos STRs ocorra de forma direcionada, gerando produtos consistentes e comparáveis entre amostras. Assim, para cada STR foram extraídas sequências de 200 pares de bases à esquerda (*upstream*) e à direita (*downstream*) do motivo repetitivo, considerando sua posição original nas amostras e a disponibilidade de sequência genômica contínua. Esse intervalo foi definido por ser suficientemente amplo para incluir regiões adequadas à ancoragem dos primers [9], [10], [11].

E. Desenho dos primers

O desenho dos oligonucleotídeos iniciadores (primers) foi realizado utilizando o software Primer3, amplamente empregado em estudos de biologia molecular para a predição de primers a partir de sequências de DNA. O Primer3 foi integrado ao pipeline por meio do pacote Python *primer3-py* (interface oficial do Primer3 para Python), adotando parâmetros amplamente recomendados na literatura para maximizar a eficiência e especificidade da reação de PCR. O tamanho dos primers foi definido entre 18 e 23 pb (ótimo: 20 pb), faixa considerada ideal para garantir especificidade adequada e eficiência de anelamento. A temperatura de melting (T_m) foi estabelecida entre 57 °C e 62 °C (ótima: 59 °C), de modo a favorecer condições homogêneas de amplificação e minimizar ligações inespecíficas. Adicionalmente, foi limitado o comprimento máximo de homopolímeros (PRIMER_MAX_POLY_X = 5), evitando regiões com repetições consecutivas de nucleotídeos que possam comprometer a estabilidade da ligação primer-template. A diferença máxima de T_m entre os pares de primers foi restringida a 2 °C, garantindo sincronização eficiente durante a etapa de anelamento da PCR [10]. A região *target* (SEQUENCE_TARGET) foi definida de modo a corresponder à porção do STR, assegurando o anelamento eficiente dos primers nas regiões adjacentes ao motivo repetitivo e, consequentemente, a amplificação específica e confiável do locus STR. A Figura 3 ilustra esquematicamente uma região de DNA contendo um STR na porção central, as regiões flangeadoras adjacentes e as posições de anelamento dos primers, bem como o amplicon teórico.

Figura 3. Localização dos primers posicionados nas regiões flangeadoras que delimitam o amplicon, que contém o motivo repetitivo (STR) entre eles.



Fonte: O autor.

F. Validação dos primers

Os amplicons teóricos – porção da sequência compreendida entre o início do *primer left* e o final do *primer reverse* – utilizados para validação *in silico* no BLASTn foram selecionados com base em critérios previamente definidos, visando garantir a especificidade, a variabilidade genética e a aplicabilidade dos marcadores propostos. Os registros foram ranqueados por meio de uma ordenação hierárquica inversa, considerando múltiplos parâmetros: (i) tamanho do STR, (ii) tamanho do amplicon e (iii) diversidade genética estimada pela entropia de Shannon [6], [12], [13]. Essa abordagem permitiu priorizar *loci* com maior potencial informativo, favorecendo aqueles com maior variabilidade e melhor adequação experimental. Finalmente, foram selecionados os 100 *loci* melhor classificados no ranqueamento para cada amostra de genoma.

Os amplicons selecionados foram então submetidos à análise de similaridade com o genoma por meio do BLASTn, com o objetivo de avaliar a especificidade das sequências amplificadas. Na busca realizada com o BLASTn, foi selecionado o banco de dados whole-genome shotgun contigs (wgs) e o organismo *Aspergillus niger* (taxid:5061).

RESULTADOS E DISCUSSÃO

A. Perfil geral dos STRs nos genomas

A triagem computacional dos genomas resultou na identificação de centenas de *loci* de STRs distribuídos ao longo dos cinco genomas analisados. A quantidade e o tipo de STRs variaram entre os genomas, refletindo o grau de diversidade intrínseca entre as cepas. A Tabela 2 apresenta parte dos resultados do levantamento inicial de STRs. Para cada cepa analisada, foi gerado um conjunto de nove arquivos padronizados conforme esse leiaute, cada um associado a uma cepa e a um tipo específico de STR.

Tabela 2. Resultado do levantamento inicial dos STRs

Unidade	Início locus	Fim locus	Cópias
A	24860975	24861058	83
T	29813117	29813183	66
AG	30838913	30838995	41
CA	4149090	4149188	49
TGT	3029117	3029381	88
CTA	1511041	1511194	51
TTA	31853043	31853160	39
TTTC	670847	670851	26
AGAA	22423545	22423609	16
ATAGG	33349853	33349973	24
GGCTCT	15156610	15156742	22
TTTCCT	20509754	20509880	21
TATGCTG	21227855	21227876	3
ATGAAAA	21845073	21845094	3
AAGAAGGA	8232364	8232500	17
TTTCTTTC	670847	670951	13
ATTTATTTT	20987328	20987616	32
GGAAACCTG	13387589	13387868	31
GTAGTAGTT	10018598	10018967	41

Descrição dos *loci* de STRs obtidos por meio do algoritmo de detecção, incluindo sua localização genômica e número de cópias. Exemplos selecionados aleatoriamente entre as cepas analisadas.

A Tabela 3 apresenta um resumo dos tipos de STR identificados por cepa. Os dados apresentados evidenciam que os STRs mono e dinucleotídicos são amplamente predominantes em todas as cepas analisadas, representando a maior fração dos motivos repetitivos identificados nos genomas, o que é coerente com a literatura que aponta maior prevalência desses tipos em organismos eucariotos [11], [12], [13]. À medida que o tamanho do motivo aumenta, observa-se uma redução progressiva no número de STRs, padrão consistente entre todas as cepas avaliadas.

As cepas analisadas exibem perfis quantitativos semelhantes, indicando uma distribuição conservada dos tipos de STR entre os genomas analisados. Pequenas variações nos valores absolutos refletem diferenças no tamanho e na composição genômica, mas não alteram

a tendência geral de decaimento do número de STRs do tipo mono- para nonanucleotídeo. Esse padrão sugere que a abundância relativa dos diferentes tipos de STR é estruturalmente conservada no genoma de *A. niger*, reforçando a adequação desses marcadores para análises comparativas e aplicações em identificação de cepas.

Tabela 3. Total de STRs por tamanho de motivo localizados em cada cepa

Tamanho do STR	CBS 513.88	CBS 554.65	KJC3	KYF3	CCTCC 206047
1	6511892	7770664	7738051	7200836	6810272
2	1118487	1349920	1346112	1251083	1181185
3	518959	619301	617747	581565	554127
4	129540	157101	156368	147672	140744
5	39817	50100	49955	47529	45476
6	18923	22911	22672	21882	21101
7	4267	5646	5717	5450	5383
8	1703	2277	2331	2202	2250
9	1394	1804	1834	1788	1752

Número absoluto de STRs classificados por tipo de motivo repetitivo (mononucleotídeos a nonanucleotídeos) identificados nos genomas das cepas analisadas.

A análise comparativa dos genomas evidenciou a coexistência de STRs conservados entre as cepas, sugerindo regiões genômicas estruturalmente estáveis, bem como a presença de STRs exclusivos, restritos a cepas específicas. Esses elementos exclusivos apresentam potencial aplicação como marcadores moleculares discriminatórios, especialmente em análises de diversidade genética e rastreamento de cepas, devido à elevada variabilidade dos microssatélites e seu alto poder de discriminação [6], [11], [12], [15], [16]. A Tabela 4 sintetiza, por cepa, o total de microssatélites (STRs), o número absoluto de *loci* exclusivos e sua representatividade percentual em relação ao total identificado.

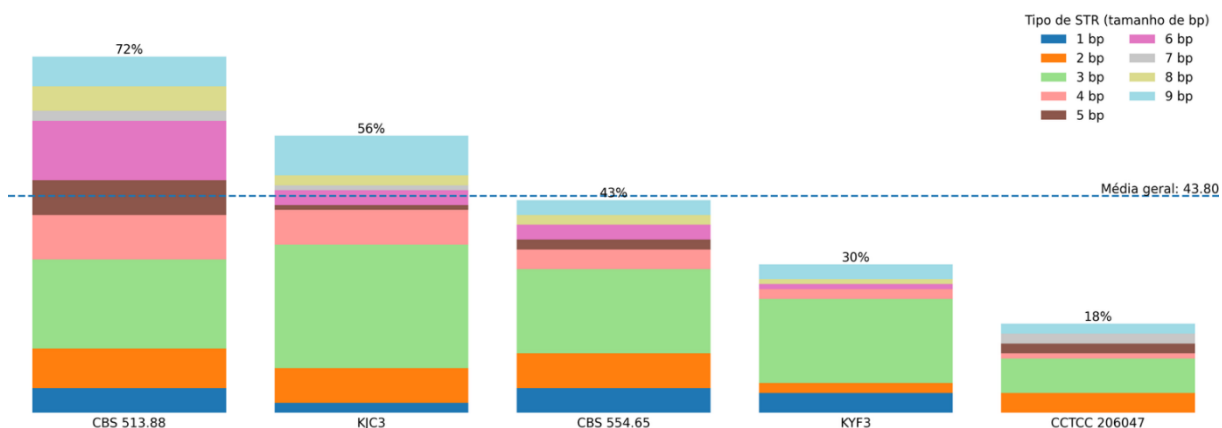
Tabela 4. Total de microssatélites (STRs) identificados por cepa de *A. niger*, incluindo o número absoluto de loci exclusivos e o respectivo percentual em relação ao total de STRs de cada genoma

Cepa	Total	Exclusivos	% Exclusivos
CBS 513.88	8344982	1047	0.0126
CBS 554.65	9979724	1133	0.0114
KJC3	9940787	918	0.0092
KYF3	9260007	485	0.0052
CCTCC 206047	8762290	774	0.0088

B. Desenho e avaliação de primers

Os primers constituem ferramentas fundamentais na biologia molecular, atuando como iniciadores em técnicas de amplificação, detecção e manipulação de ácidos nucleicos, como a reação em cadeia da polimerase (PCR). A utilização do Primer3 via script Python permitiu automatizar a análise, garantindo a seleção de oligonucleotídeos com baixo risco de formação de estruturas secundárias ou dímeros. A validação desses primers foi realizada por meio de análises *in silico* utilizando o BLASTn, a partir das sequências correspondentes aos produtos de amplificação teóricos. Essas sequências foram extraídas da sequência genômica original com base nas extremidades definidas pelo alinhamento dos primers, de modo a representar com precisão os amplicons teóricos gerados pela reação de PCR. Dessa forma, a análise por BLASTn confirmou a elevada especificidade de uma parcela significativa dos primers avaliados, com média de 43,80% de primers validados, variando entre 18% e 72% entre os conjuntos analisados (Figura 4). Esses primers apresentaram alinhamentos únicos ou preferenciais no genoma da cepa alvo de *A. niger*, associados a valores de *e-value* extremamente baixos, próximos de zero, indicando alta confiabilidade na identificação da cepa alvo e reduzindo substancialmente a probabilidade de alinhamentos inespecíficos em outros genomas. Essa etapa de validação é fundamental para assegurar que os primers apresentem desempenho adequado em aplicações experimentais subsequentes, como ensaios de PCR convencional, minimizando o risco de amplificações cruzadas ou não específicas.

Figura 4. Percentuais de amplicons com sucesso na identificação da cepa correta. Os resultados de BLASTn que atenderam ao filtro de único resultado com 100% de alinhamentos.



Fonte: O autor.

Para cada cepa analisada, foi elaborada uma tabela contendo os *loci* STR selecionados que atenderam aos critérios de filtragem dos resultados do BLASTn, apresentando alinhamento único e 100% de identidade. As tabelas disponíveis em Material Suplementar 1 apresentam os dados referentes aos primers projetados para cada locus STR, bem como as sequências dos amplicons, assegurando a reprodutibilidade experimental e viabilizando a aplicação desses *loci* como marcadores genéticos com elevado poder discriminatório entre cepas de *A. niger*.

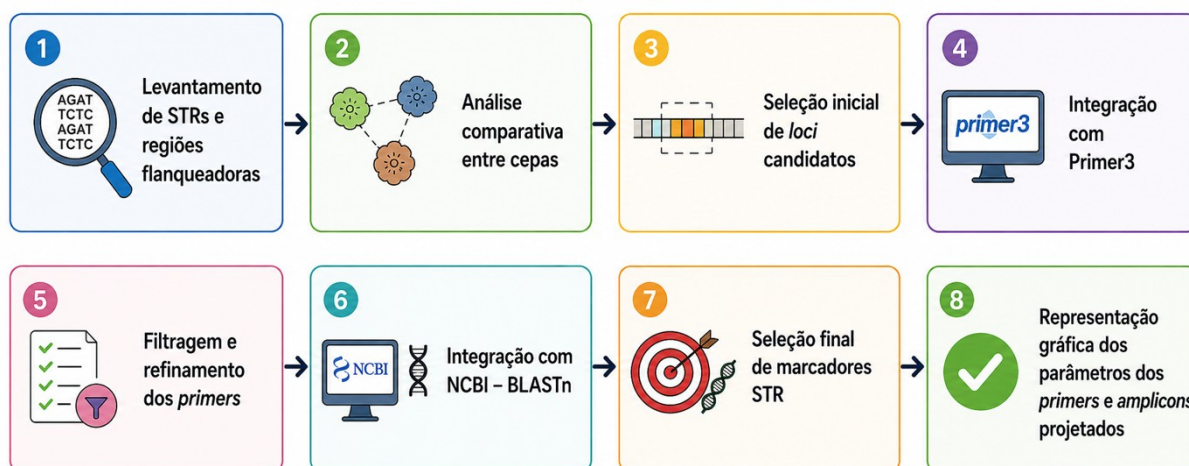
C. Seleção final dos marcadores

Os amplicons que apresentaram alinhamento bem-sucedido, foram submetidos a uma nova etapa de ranqueamento por ordenação hierárquica inversa, com base no tamanho do STR, no tamanho do amplicon e na diversidade genética (entropia) das amostras. Para cada cepa, os dois *loci* mais bem classificados foram selecionados como marcadores candidatos ideais para sua identificação.

D. Visão geral do pipeline computacional integrado

A Figura 5 ilustra as etapas do pipeline computacional integrado desenvolvido neste estudo para a identificação, caracterização e seleção racional de marcadores STR com potencial aplicação em sistemas de rastreabilidade de bioinsumos agrícolas.

Figura 5. Representação gráfica dos passos do pipeline computacional.



Fonte: O autor.

E. Potencial aplicabilidade dos STRs identificados

Os resultados obtidos neste estudo destacam o potencial dos STRs como marcadores moleculares em cepas de *A. niger*. Os *loci* identificados, especialmente os exclusivos de uma cepa ou com alta entropia, podem ser explorados em estudos de tipagem molecular, monitoramento de cepas de bioinsumos em todas as etapas de produção e após aplicação no campo, rastreabilidade de contaminantes ou bioensaios de diversidade microbiana. A abordagem computacional *in silico* demonstrou-se eficiente e reproduzível, constituindo uma alternativa econômica e escalável para a triagem inicial de *loci* candidatos antes da validação experimental em laboratório.

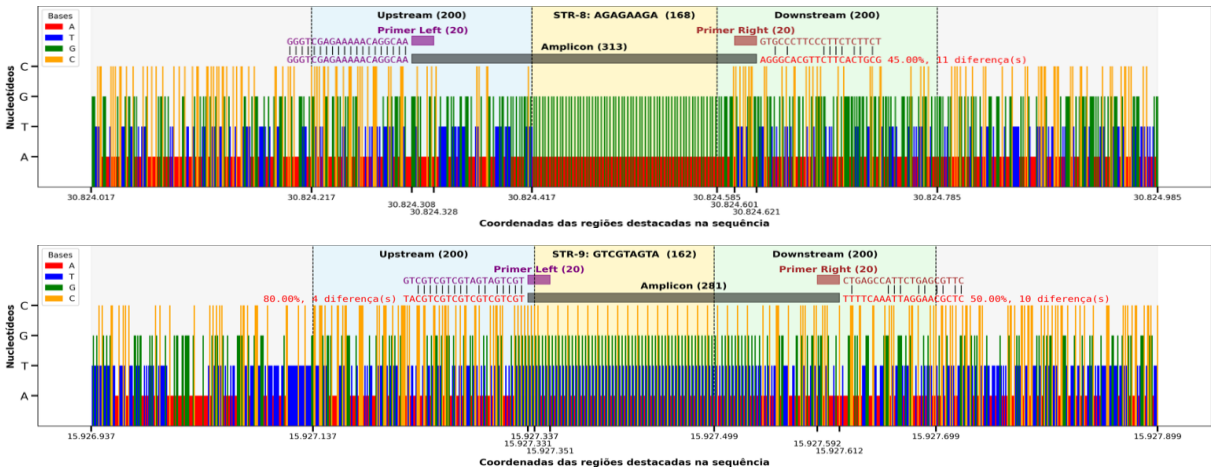
F. Disponibilidade de dados e código fonte

Os scripts desenvolvidos em Python foram disponibilizados na forma de *notebooks*, juntamente com os conjuntos de dados utilizados nas análises e as imagens geradas ao longo do estudo. O código-fonte e os materiais suplementares estão disponíveis publicamente na plataforma GitHub, no endereço: https://github.com/warioiraw/analise_str.git. O código foi desenvolvido utilizando bibliotecas amplamente empregadas em bioinformática e análise de dados em Python, cujas versões e dependências estão especificadas no próprio repositório. Todo o material é disponibilizado sob licença aberta, permitindo sua reutilização, modificação e adaptação para fins acadêmicos, desde que devidamente citada a fonte original. Essa iniciativa visa promover a transparência metodológica, facilitar a validação independente dos resultados e incentivar a aplicação do pipeline computacional em estudos futuros envolvendo a análise de microssatélites em genomas fúngicos.

G. Representação gráfica dos parâmetros dos primers e amplicons projetados

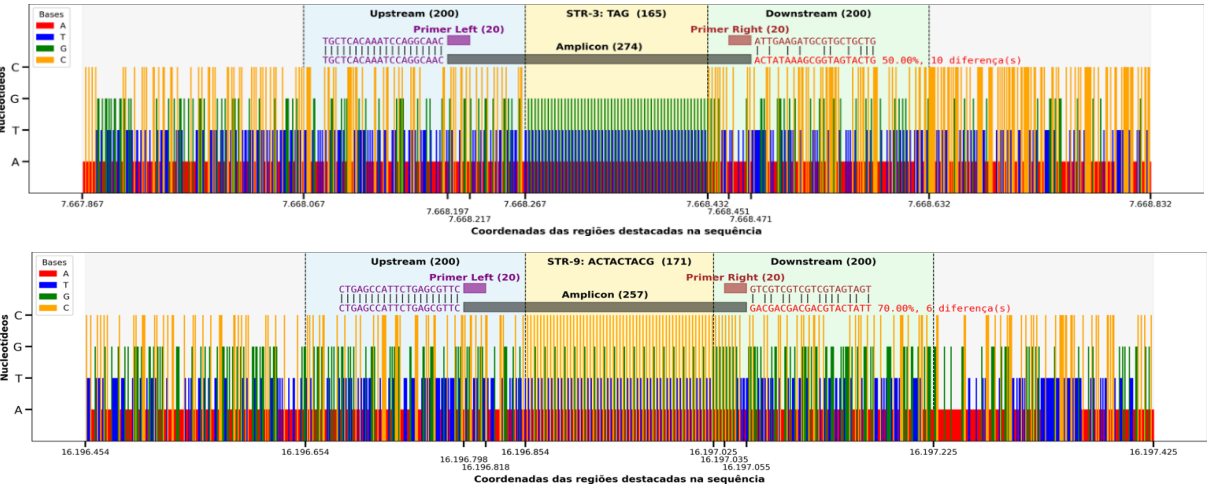
As figuras 6 a 10 foram geradas a partir desses resultados com o objetivo de auxiliar na interpretação por meio de representações gráficas da distribuição, variabilidade e características dos *loci* STR identificados. Essas visualizações permitem uma compreensão mais intuitiva dos padrões genômicos observados, contribuindo para a avaliação dos mesmos.

Figura 6. Representação gráfica dos parâmetros dos primers e amplicons projetados para a cepa CBS 513.8. Topo: STR-8 AGAGAATA. Base: STR-9 GTCGTAGTA.



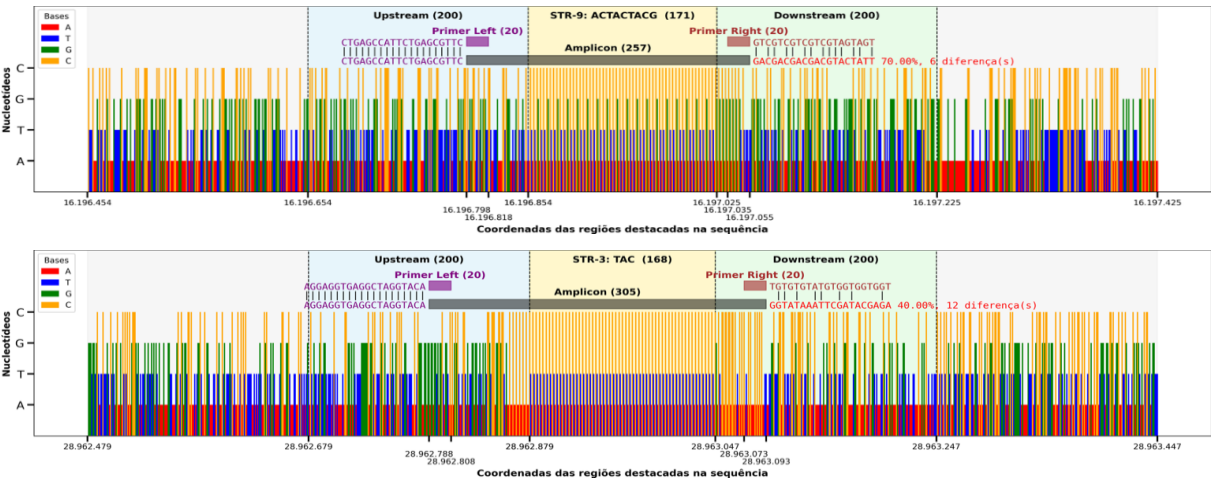
Fonte: O autor.

Figura 7. Representação gráfica dos parâmetros dos primers e amplicons projetados para a cepa CBS 554.65. Topo: STR-3 TAG. Base: STR-9 ACTACTACG.



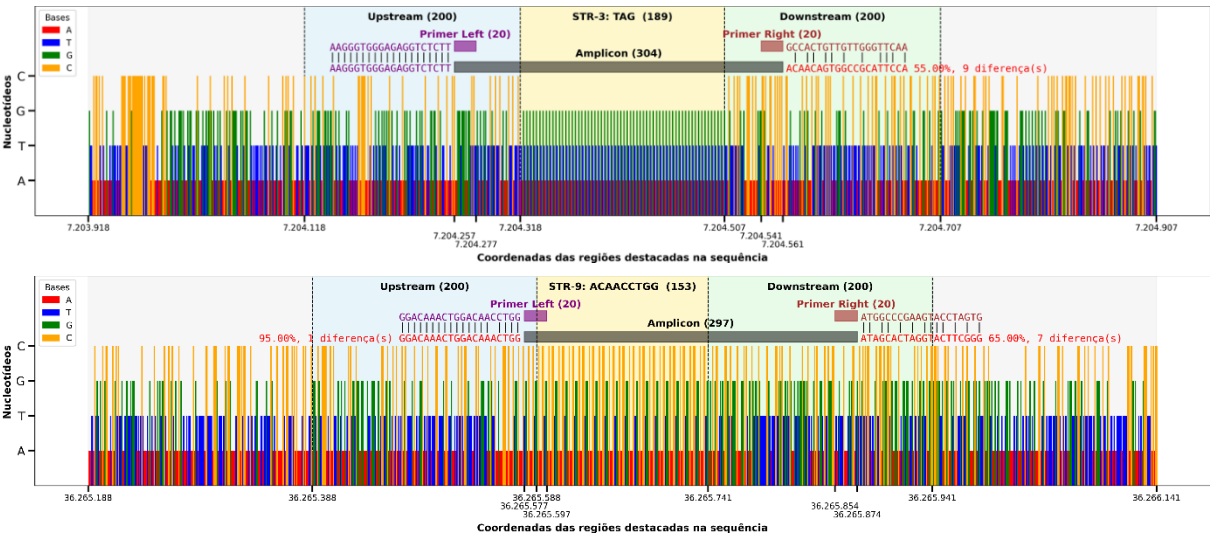
Fonte: O autor.

Figura 8. Representação gráfica dos parâmetros dos primers e amplicon projetados para a cepa CBS KJC3. Topo: STR-9 ACTACTACG. Base: STR-3 TAC.



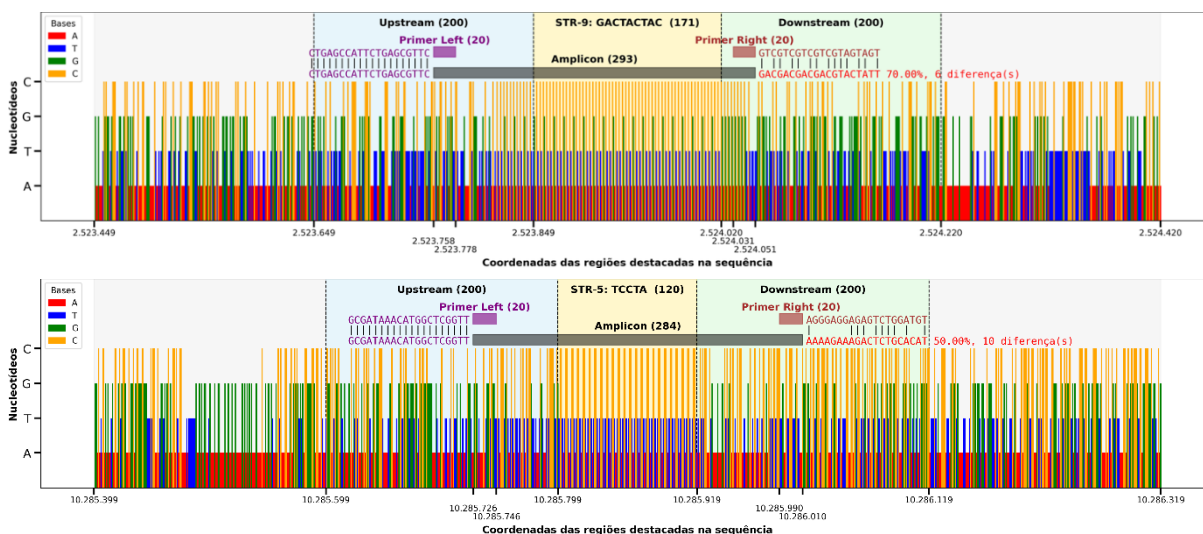
Fonte: O autor.

Figura 9. Representação gráfica dos parâmetros dos primers e amplicon projetados para a cepa KYF3. Topo: STR-3 TAG. Base: STR-9 ACAACCTGG.



Fonte: O autor

Figura 10. Representação gráfica dos parâmetros dos primers e amplicon projetados para a cepa CCTCC 206047. Topo: STR-9 GACTACTAC. Base: STR-5 TCCTA.



Fonte: O autor.

Como próximos passos, propõe-se a validação experimental dos microssatélites classificados como ideais por meio de técnicas de biologia molecular, visando confirmar a especificidade, sensibilidade e reprodutibilidade dos marcadores desenvolvidos. Inicialmente, os primers desenhados serão avaliados em reações de PCR utilizando DNA genômico puro das cepas-alvo, a fim de verificar a amplificação esperada e o tamanho dos amplicons previstos *in silico*. Posteriormente, será realizada a avaliação da especificidade dos marcadores por meio de testes cruzados com isolados filogeneticamente próximos e outros microrganismos presentes no mesmo ambiente, permitindo identificar possíveis reações inespecíficas. Após a validação individual dos *loci*, será investigada a viabilidade da combinação de múltiplos marcadores em reações de PCR multiplex, visando aumentar a eficiência e reduzir custos analíticos. Por fim, pretende-se validar os marcadores em condições ambientalmente mais próximas da aplicação prática, por meio da detecção das cepas em amostras de solo artificialmente inoculadas. Esses experimentos permitirão avaliar a capacidade dos marcadores em detectar e monitorar as cepas-alvo em matrizes complexas, contribuindo para futuras aplicações em rastreamento e monitoramento ambiental de bioinoculantes.

CONCLUSÕES

Esse trabalho desenvolveu um pipeline computacional integrado, robusto e customizado para a identificação, caracterização e validação *in silico* de microssatélites (STRs) no genoma

do fungo *A. niger*. Utilizando um algoritmo próprio foi possível detectar STRs, eliminar *loci* redundantes, quantificar sua variabilidade e, por fim, propor primers específicos para diferenciação de cepas de *A. niger*.

A aplicação de métricas informativas, como o comprimento das repetições e a Entropia de Shannon, possibilitou a seleção racional dos *loci* mais promissores, priorizando regiões com maior variabilidade e, portanto, maior valor como marcadores moleculares. O desenho de primers a partir das regiões flanqueadoras, seguido de avaliação via Primer3 e validação por BLASTn, assegurou a especificidade e viabilidade dos oligonucleotídeos gerados. A seleção final dos primers submetidos à verificação por BLASTn foi conduzida com base em critérios mínimos previamente estabelecidos, os quais podem ser ajustados de acordo com os objetivos específicos da análise. Tais ajustes incluem, por exemplo, a expansão ou redução do tamanho das regiões flanqueadoras, visando a otimização do posicionamento dos primers, processo que pode ser avaliado visualmente por meio das figuras geradas a partir dos resultados obtidos.

De forma geral, os resultados indicam que abordagens computacionais sistemáticas são ferramentas promissoras para acelerar a descoberta de STRs informativos e o desenvolvimento de marcadores específicos para cepas, contribuindo para estratégias mais precisas de caracterização molecular em *A. niger* e outros organismos de interesse biotecnológico, reforçando papel estratégico de ferramentas bioinformáticas no avanço da genômica aplicada à agricultura. Em particular, o conhecimento detalhado dos STRs de *A. niger* abre perspectivas para a rastreabilidade genética de cepas de bioinsumos. Além disso, o pipeline proposto neste trabalho pode ser adaptado para a análise de outros fungos e bactérias de interesse agrícola, incluindo microrganismos amplamente empregados como bioinsumos, tais como espécies dos gêneros *Bacillus*, *Pseudomonas*, *Azospirillum* e *Trichoderma*. Como a metodologia é baseada em dados genômicos e composta por etapas modulares, sua reutilização requer, em grande parte, apenas a substituição dos genomas de entrada e o ajuste de parâmetros específicos, como o tamanho dos motivos STR e critérios de seleção dos *loci*.

REFERÊNCIAS

- [1] MENDES, R.; GARBEVA, P.; RAAIJMAKERS, J. M. The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiology Reviews*, v. 37, n. 5, p. 634–663, 2013. <https://doi.org/10.1111/1574-6976.12028>
- [2] BANO, S.; WU, X.; ZHANG, X. Towards sustainable agriculture: rhizosphere microbiome engineering. *Applied Microbiology and Biotechnology*, v. 105, n. 19, p. 7141–7160, 2021. <https://doi.org/10.1007/s00253-021-11555-w>
- [3] VASSILEVA, M. et al. Fungi, P-solubilization, and plant nutrition. *Microorganisms*, v. 10, n. 9, p. 1716, 2022. <https://doi.org/10.3390/microorganisms10091716>
- [4] MUNDIM, G. S. M.; MACIEL, G. M.; MENDES, G. O. *Aspergillus niger* as a biological input for improving vegetable seedling production. *Microorganisms*, v. 10, n. 4, p. 674, 2022. <https://doi.org/10.3390/microorganisms10040674>
- [5] MANFREDINI, A. et al. Current Methods, Common Practices, and Perspectives in Tracking and Monitoring Bioinoculants in Soil. *Frontiers in Microbiology*, v. 12, p. 698491, 2021. <https://doi.org/10.3389/fmicb.2021.698491>
- [6] SPRUIJTENBURG, B.; MEIJER, E. F. J. Short tandem repeat genotyping of medically important fungi: A comprehensive review of a powerful tool with extensive future potential. *Mycopathologia*, v. 189, 2024. <https://doi.org/10.1007/s11046-024-00877-8>
- [7] MCCOUCH, S. R. et al. Microsatellite marker development, mapping and applications in rice genetics and breeding. *Plant Molecular Biology*, v. 35, p. 89–99, 1997. <https://doi.org/10.1023/A:1005711431474>
- [8] TEMNYKH, S. et al. Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). *Theoretical and Applied Genetics*, v. 100, p. 697–712, 2000. <https://doi.org/10.1007/s001220051342>
- [9] SAMBROOK, J.; RUSSELL, D. W. *Molecular Cloning: A Laboratory Manual*. The Quarterly Review of Biology, v. 76, n. 3, 2001. <https://doi.org/10.1086/394015>
- [10] UNTERGASSER, A. et al. Primer3—new capabilities and interfaces. *Nucleic Acids Research*, v. 40, n. 15, p. e115, 2012. <https://doi.org/10.1093/nar/gks596>
- [11] TAUTZ, D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research*, v. 17, n. 16, p. 6463–6471, 1989. <https://doi.org/10.1093/nar/17.16.6463>
- [12] ELLEGREN, H. Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics*, v. 5, p. 435–445, 2004. <https://doi.org/10.1038/nrg1348>
- [13] TÓTH, G.; GÁSPARI, Z.; JURKA, J. Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Research*, v. 10, n. 7, p. 967–981, 2000. <https://doi.org/10.1101/gr.10.7.967>
- [14] HAMMING, W. Error detecting and error correcting codes. *Bell System Technical Journal*, v. 29, n. 2, p. 147–160, 1950. <https://doi.org/10.1002/j.1538-7305.1950.tb00463.x>
- [15] ZHU, Y. et al. SSR molecular marker developments and genetic diversity analysis of *Zanthoxylum nitidum* (Roxb.) DC. *Scientific Reports*, v. 9, p. 19817, 2019. <https://doi.org/10.1038/s41598-023-48022-7>
- [16] FAN, H.; CHU, J. A brief review of short tandem repeat mutation. *Genomics, Proteomics & Bioinformatics*, v. 5, n. 1, p. 7–14, 2007. [https://doi.org/10.1016/s1672-0229\(07\)60009-6](https://doi.org/10.1016/s1672-0229(07)60009-6)

[illegible]

[illegible]

Tabela Suplementar 2. CEPA CBS 554.65

[illegible]

[illegible]

G	32	CGGGATGGTAAAGGAATC CAA	ATCTTCCCCGACTA GGTCT	317
TTTC	8	GACATGATGACGAGCGC TTT	TGCTAGTTCGATGA GTCC	212
TTG	10	GTGGATTITGCCATGGT GGT	TCCGAACCTCCTCAT CTCCA	271
CTGCC	6	TCCCCTTTTGCCAAGAT CT	CGGACCCTGACTCT ATGAG	260
GAC	10	CAGGTTTCAACGGCCAT CAA	TTCTCTCTCTTGCTC TGGC	207
CTGCTA	5	GCATCTCCCCTATTGCT GC	ACCACGAGAGATGCT TGGTT	187

[illegible]

[illegible][illegible]

[illegible]

Material Suplementar 2. Pipeline Computacional – fluxogramas do processo

Os algoritmos foram desenvolvidos seguindo uma abordagem modular, na qual cada etapa exporta seus resultados em arquivos CSV que servem como entrada para a etapa subsequente. Essa estratégia permite retomar o processamento a partir de pontos específicos, facilita a reprodutibilidade e atua como um registro estruturado de cada fase do pipeline.

A. *Triagem Inicial dos STRs*

A Figura 11 ilustra em duas etapas o processo inicial de levantamento dos STRs.

A) Identificação de STRs

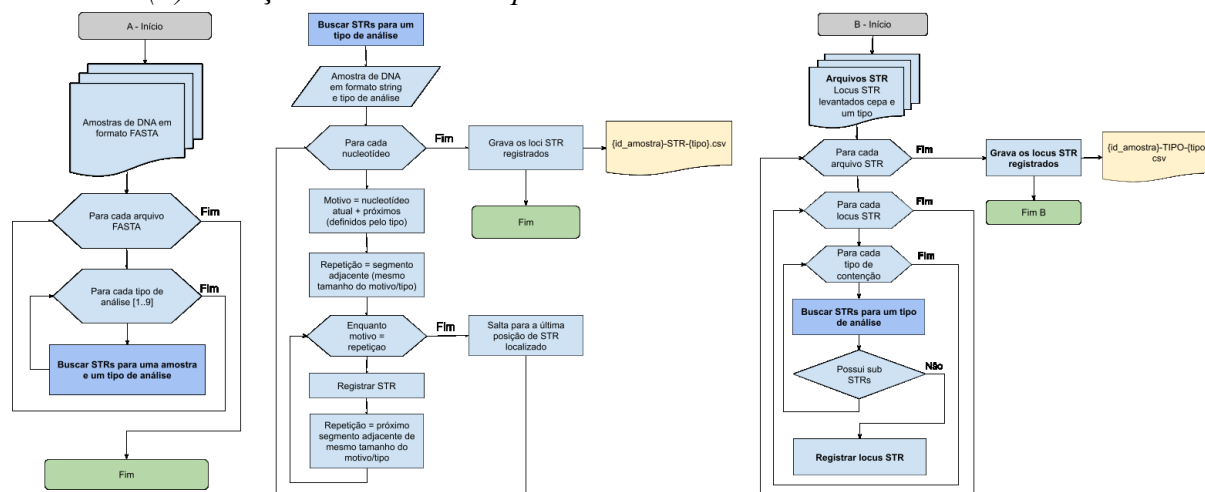
Esta etapa consiste na detecção sistemática de sequências repetitivas em tandem (STRs) em genomas completos. Para cada sequência proveniente de um arquivo FASTA — correspondente a uma cepa analisada — é realizada uma varredura sequencial em busca de repetições cujo motivo possui comprimento variável de 1 a 9 nucleotídeos (SPRUIJTENBURG; MEIJER, 2024). O processo de detecção foi encapsulado em uma função independente, desacoplada da leitura das amostras, visando à reutilização na próxima fase. Como saída, uma lista de loci STR identificados é exportada em formato CSV.

B) Eliminação de STRs Redundantes

Após a detecção inicial, foi realizada a filtragem de STRs redundantes, com ênfase nos casos de “STRs contidos em STRs”, nos quais motivos repetitivos de menor unidade estão completamente embutidos e reiterados em STRs de motivo mais longo. A identificação dessas redundâncias baseia-se em uma tabela de contenção, que descreve relações de equivalência e ambiguidade entre STRs de diferentes tamanhos de motivo. Essa tabela estabelece quais motivos curtos podem gerar repetições estruturalmente indistinguíveis de motivos mais longos. As relações consideradas neste estudo foram:

- STRs de motivo 9 podem apresentar ambiguidade com 3 e 1;
- STRs de motivo 8 com 4, 2 e 1;
- STRs de motivo 6 com 3, 2 e 1;
- STRs de motivo 4 com 2 e 1;
- os demais tipos apresentam ambiguidade exclusivamente com STRs de motivo 1.

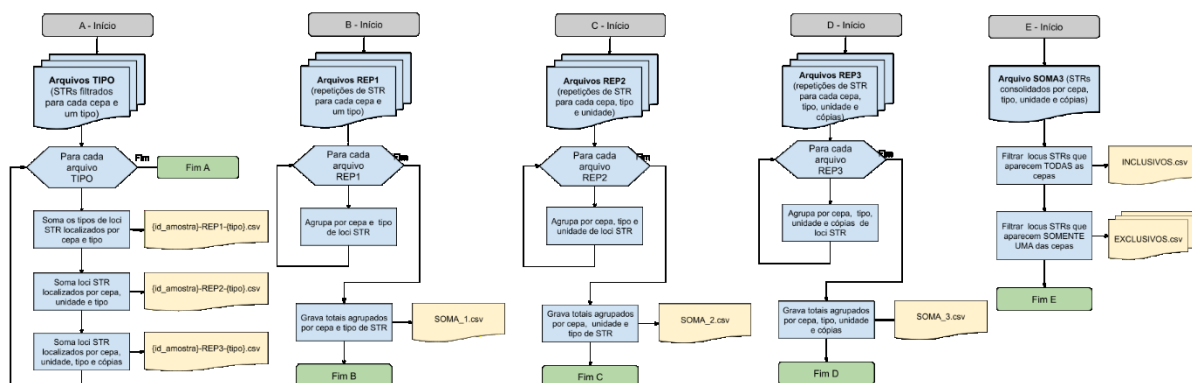
Figura 11. *Detecção - composto por duas etapas principais: (A) detecção sequencial dos loci STR e (B) remoção dos STRs sobrepostos.*



B. Agregação, Quantificação e Classificação de STRs

Conforme ilustrado na Figura 12, esta etapa considera múltiplos níveis de organização. Inicialmente é realizada a soma do número total de STRs por cepa analisada, bem como a quantificação estratificada por cepa e tipo de STR e por cepa, tipo e tamanho (número de cópias). Em seguida, os dados são agrupados entre todas as cepas, consolidando-se as quantificações por tipo de STR e tamanho, de modo a obter o número de STRs localizados para cada combinação de motivo e número de cópias em cada cepa. A partir desse agrupamento global, os loci são então classificados conforme seu padrão de ocorrência entre as amostras. Foram identificados como STRs inclusivos aqueles loci que apresentam motivos e tamanhos equivalentes (mesmo número de cópias) e que ocorrem em todas as sequências analisadas. Em contraste, são definidos como STRs exclusivos os loci que apresentam motivos e tamanhos equivalentes, mas que ocorrem em apenas uma das cepas analisadas. Foram gerados arquivos distintos contendo os STRs inclusivos e os STRs exclusivos, permitindo a análise comparativa da conservação e da especificidade dos loci STR entre as cepas estudadas.

Figura 12. Etapas sucessivas de agregação por cepa, tipo, unidade e número de cópias (REP1, REP2 e REP3). As informações consolidadas são então combinadas em um arquivo resumo global (SOMA3), a partir do qual são identificados os STRs inclusivos, presentes em todas as cepas analisadas, e os STRs exclusivos, presentes em apenas uma cepa, considerando motivos e tamanhos equivalentes.



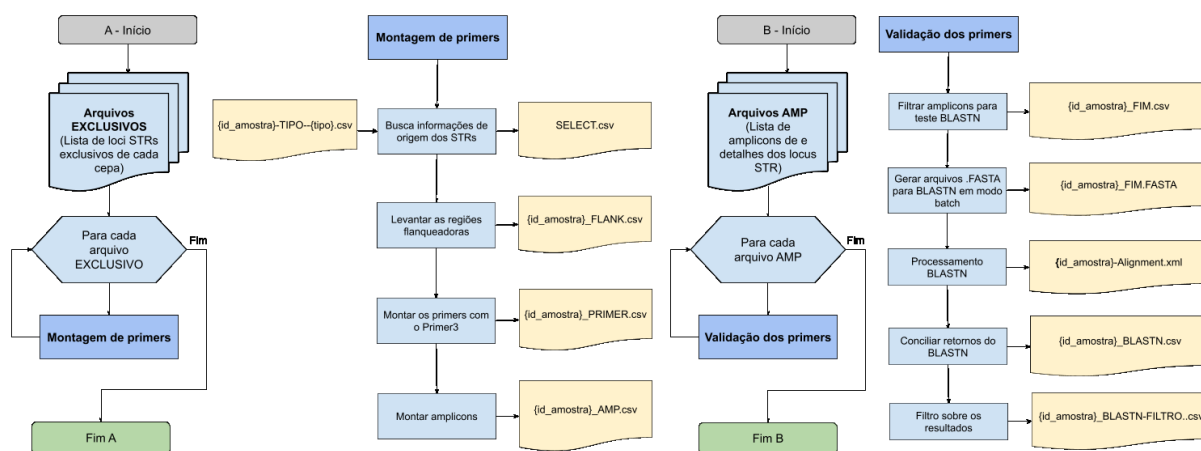
Fonte: O autor

C. Montagem e Validação dos Primers

A Figura 13 detalha o processo de montagem e validação dos primers. Para cada STR exclusivo identificado, foram recuperadas as informações de origem — incluindo posição genômica e características do motivo repetitivo — a partir dos arquivos `id_cepas-tipo-X.csv`, gerados previamente na etapa de detecção dos STRs. Em seguida, para cada locus, foram extraídas as sequências flanqueadoras (upstream e downstream) que, juntamente com o STR central, compõem a sequência utilizada como entrada para o Primer3. O Primer3 é então acionado individualmente para cada locus, retornando até cinco conjuntos possíveis de primers. A partir dessas informações, os amplicons são reconstruídos considerando as posições exatas de ancoragem dos primers na sequência analisada. Os conjuntos de primers selecionados têm seus amplicons submetidos à validação por BLASTn, com objetivo de confirmar a especificidade, descartar alinhamentos inesperados e garantir que cada par amplifique apenas o locus desejado em cada cepa. Após a validação, foi executada uma etapa de conciliação dos resultados, na qual são integradas as informações de desenho, posição e especificidade dos primers. Foram considerados como critérios de validação a cobertura da sequência de consulta (query cover) e o percentual de identidade (percent identity). A cobertura da consulta foi interpretada como indicativa do tamanho efetivo do amplicon alinhado, sendo priorizados os resultados que apresentaram query cover correspondente à totalidade do amplicon teórico, assegurando que todo o amplicon estivesse representado no alinhamento. Adicionalmente, foi exigido que os alinhamentos apresentassem 100% de identidade nucleotídica exclusivamente com a cepa-alvo de *A. niger*. Esse critério foi adotado com o objetivo de garantir elevada

especificidade dos produtos teóricos de PCR, evitando amplificações cruzadas com outras cepas ou organismos filogeneticamente relacionados. O resultado é um conjunto final de arquivos, organizados por cepa, contendo os STRs efetivamente discriminatórios e seus primers validados.

Figura 13. Fluxo de montagem e validação dos primers. (A) Recuperar as informações de origem, extrair as regiões flanqueadoras que, acionar o Primer3. Montar os amplicons para cada conjunto de primers. (B) Os amplicons são submetidos à validação por BLASTN para verificação de especificidade em seguida é feita a conciliação dos resultados do BLASTN e aplicação de filtros de especificidade.



Fonte: O autor