VARIANTES ESTRUTURAIS EM GENES DE *Coffea canephora* (CONILON) IDENTIFICADAS POR SEQUENCIAMENTO NANOPORE

Mariana Cansian Sattler, Guilherme Cordenonsi da Fonseca, Rosana Gomes de Oliveira, Adésio Ferreira, Marcia Flores da Silva Ferreira

Universidade Federal do Espírito Santo (UFES), Departamento de Agronomia, Alto Universitário, s/nº, Bairro Guararema, Alegre – ES, CEP 29500-000, Brasil, marianacansattle@gmail.com.

Resumo

Coffea canephora é a principal espécie de café cultivada no Espírito Santo, mas sua diversidade genética tem sido avaliada quase exclusivamente por SNPs e SSRs, permanecendo as variantes estruturais (SVs) pouco exploradas. Este trabalho identificou SVs divergentes entre o clone 12V da cultivar Vitória e o genoma de referência da espécie, a partir de uma única corrida de sequenciamento por nanoporos, cujas leituras se alinharam de forma colinear à referência. Após a aplicação de filtros sucessivos, como filtros de borda de *contig*, confirmação por leituras e interseção com a anotação da referência, foram identificados 156 pares gene-SV em 147 genes, dos quais oito apresentaram ponto de quebra interno à sequência codificadora. Além disso, um gene de resistência do tipo NBS-LRR apresentou inserção de 53 pb na região 5′-UTR. Os resultados evidenciam que uma única corrida de sequenciamento por *long-reads* ancorada a um genoma de referência tem o potencial de localizar divergências estruturais em genes específicos, indicando alvos candidatos para validação experimental e aplicação em programas de melhoramento genético.

Palavras-chave: Café. Variantes estruturais. *Long-reads*. Genômica.

Área do Conhecimento: Ciências Agrárias – Agronomia – Melhoramento Vegetal.

Introdução

O café é uma das bebidas mais consumidas no mundo, mas sua produtividade está ameaçada pelas mudanças climáticas, sobretudo por secas prolongadas e temperaturas elevadas (Jawo et al., 2023). No Espírito Santo, mais de 65% da produção corresponde a *Coffea canephora*, o que torna o estado o principal produtor dessa espécie no Brasil (Baqueta et al., 2024). A espécie se destaca pela elevada produtividade, resistência a doenças e adaptabilidade a climas quentes (Ferrão et al., 2017). Além disso, por ser alógama e autoincompatível, exibe elevados níveis de diversidade genética e fenotípica em populações nativas e cultivadas (Cubry et al., 2013; Zaidan et al., 2023; De Oliveira et al., 2026).

A diversidade genética de *C. canephora* tem sido avaliada predominantemente por Single Nucleotide Polymorphisms (SNPs) e Simple Sequence Repeats (SSRs), enquanto as variantes estruturais (SVs) permanecem pouco exploradas. As SVs respondem por uma parcela relevante da diversidade funcional, mas não são facilmente detectadas por abordagens de sequenciamento baseadas em leituras curtas (Yuan et al., 2021). Famílias gênicas ligadas à defesa contra patógenos e ao estresse, as quais ocorrem com frequência em múltiplas cópias organizadas em tandem, estão sujeitas a duplicações e rearranjos estruturais (Grover; Wendel, 2010). Em *Medicago truncatula*, a família NBS-LRR está entre as mais enriquecidas na fração acesso-específica do genoma que possui SVs (Zhou et al., 2017), e em *Glycine max* as SVs concentram-se em *clusters* de genes de resposta a estresse biótico (Liu et al., 2020). Nesse contexto, o sequenciamento de *long-reads*, mesmo em coberturas baixas e ancorado a um genoma de referência, está entre as estratégias mais eficientes para a detecção das SVs (Smolka et al., 2024; Polkhovskaya et al., 2025).

Estão disponíveis para *C. canephora* o genoma do acesso duplo-haploide DH200-94 (Denoeud et al., 2014) e sua remontagem mais recente (Salojärvi et al., 2024), ambos disponibilizados pelo Coffee Genome Hub (https://coffee-genome-hub.southgreen.fr/; Dereeper et al., 2015). Embora uma única corrida de sequenciamento não seja suficiente para a montagem de um genoma de referência, a ancoragem de uma corrida de *long-reads* contra um genoma de referência pode possibilitar a localização de divergências estruturais em genes específicos. Desse modo, o objetivo deste trabalho foi identificar variantes estruturais divergentes entre um genótipo representativo do “Conilon Capixaba” e

o genoma de referência de *C. canephora*, a partir de dados preliminares de uma única corrida de sequenciamento por nanoporos.

Metodologia

Amostras de folhas jovens do clone 12V (cultivar Vitória Incaper 8142) de *C. canephora* foram coletadas e submetidas à extração de DNA por meio do protocolo CTAB (Doyle; Doyle, 1990) com modificações. A qualidade e quantidade do DNA foram avaliadas por meio de eletroforese em gel de agarose a 0,8%, Nanodrop e Qubit. A biblioteca genômica foi construída utilizando o kit Ligation Sequencing V14, compatível com flow cells R10.4.1 (ONT). O sequenciamento foi conduzido em plataforma MinION Mk1B (flow cells R10.4.1, ONT) com basecalling posterior (Dorado).

As leituras brutas foram avaliadas com NanoPlot (De Coster et al., 2018) e filtradas com chopper (qualidade média ≥ Q10; comprimento ≥ 500 pb). A montagem *de novo* foi realizada com Flye (Kolmogorov et al., 2019), a partir do subconjunto de leituras Q15 com comprimento ≥ 1 kb, e sua completude foi avaliada com BUSCO (Manni et al., 2021). A sintenia global com o genoma de referência de *C. canephora* foi avaliada por *dot plot* no D-GENIES (Cabanettes; Klopp, 2018).

Para a anotação foi utilizada como referência a nova versão do genoma do acesso DH200-94 de *C. canephora* (GCA_036785865.1; Salojärvi et al., 2024) e seus arquivos de anotação disponibilizados no Coffee Genome Hub. Os dados da anotação foram transferidos para os *contigs* da montagem de Conilon 12V com o Liftoff (Shumate; Salzberg, 2021), adotando-se cobertura de alinhamento ≥ 0,95 como critério de gene completo. As leituras filtradas foram alinhadas à referência com o minimap2 (Li, 2018). Os alinhamentos foram ordenados e indexados com o SAMtools (Danecek et al., 2021) e a profundidade por gene foi calculada com o mosdepth (Pedersen; Quinlan, 2018).

Para a detecção de SVs os *contigs* foram alinhados à referência com o minimap2 (preset asm20) e as SVs com tamanho ≥ 50 pb foram chamadas com o SVIM-asm (Heller; Vingron, 2020). Sobre a chamada bruta aplicaram-se filtros sucessivos com o BEDTools (Quinlan; Hall, 2010), incluindo a remoção de SVs situadas a menos de 2 kb das extremidades de *contig*, a confirmação independente pelas variantes chamadas diretamente nas leituras com o Sniffles2 (Smolka et al., 2024), e a interseção com os modelos gênicos e sequências codificadoras. Definiu-se como conjunto de maior confiança o subconjunto de pares situados no núcleo do *contig* e com maior número de leituras de suporte. Cada par gene-SV foi então confrontado com o atributo *coverage* da anotação transferida por Liftoff, excluindo-se os eventos que contêm o gene integralmente sem ponto de quebra interno à sequência codificadora, como deleções sobre genes recuperados com cobertura ≥ 0,95 e inversões cujos pontos de quebra situam-se fora do modelo gênico.

Resultados

A corrida de sequenciamento produziu 7,34 milhões de leituras, totalizando 5,38 Gb, com N50 de 1.161 pb, comprimento médio de 427 pb e qualidade de Q14,6. Após a filtragem restaram 2,82 milhões de leituras (3,61 Gb), com comprimento médio de 1.278 pb, preservando cerca de 67% das bases e resultando em cobertura de aproximadamente 5,1× sobre um tamanho de genoma estimado em 710 Mb. A montagem *de novo* obtida nessa cobertura é fragmentada (BUSCO com 6,9% de ortólogos completos), razão pela qual a análise foi conduzida por ancoragem ao genoma de referência, e não por comparação entre montagens *de novo*.

Do total de leituras filtradas, 97,70% produziram alinhamento contra a referência, confirmando a identidade da amostra como *C. canephora*. A profundidade média situou-se entre 4× e 5× nos cromossomos. O *dot plot* global revelou uma diagonal contínua e colinear, sem evidências de rearranjos de larga escala entre o clone 12V de Conilon e a referência. A transferência de anotação por meio do Liftoff recuperou 8.429 genes, dos quais 2.777 atenderam ao critério de completude. A cobertura por gene identificou 18.386 genes com profundidade ≥ 3×, 9.193 com ≥ 5× e 1.030 com ≥ 10×.

A chamada bruta de variantes produziu 6.298 SVs (2.794 deleções, 2.498 inserções, 101 inversões, 26 duplicações e 879 *breakends*), totalizando 7.177 registros no arquivo de variantes, uma vez que cada *breakend* é registrado em suas duas extremidades. O descarte de variantes situadas a menos de 2 kb das extremidades dos *contigs* eliminou cerca de 75% desse conjunto, e a exigência de suporte concordante nas leituras reduziu-o a 1.599 registros. A interseção com regiões gênicas identificou 156 pares gene-SV em 147 genes, dos quais 80 pares, em 76 genes, reuniram simultaneamente

posição no núcleo do *contig* e maior número de leituras de suporte. Entre os 156 pares gene-SV predominaram inserções (72) e deleções (62), seguidas de *breakends* (12), inversões (6) e duplicações em tandem (4), proporção invertida em relação à chamada bruta, na qual as deleções eram a classe mais numerosa. Quanto à classificação funcional, 109 pares corresponderam a genes com função putativa, 28 a proteínas hipotéticas, 10 a genes funcionalmente anotados e 9 a sequências codificadoras derivadas de transposons.

A confrontação desse conjunto com a anotação transferida pelo Liftoff revelou falsos positivos remanescentes: entre os 147 genes, 73 pares gene-deleção se sobrepõem e, em oito genes, a deleção cobre o gene por completo, mas o Liftoff localizou o gene na montagem com cobertura ≥ 0,95, de modo que nenhum gene inteiramente contido na deleção está ausente da montagem. O caso mais extenso é a deleção de 13.552 pb em Chr02:26.459.285–26.472.837, que abrangeria três genes vizinhos, Ccan002g032890, Ccan002g032900 e Ccan002g032910. Os três genes estão no *contig_4072*, com cobertura de 0,998, 1,0 e 1,0, respectivamente, na mesma ordem e espaçamentos da referência. O *contig* que carrega o bloco situa-se inteiramente dentro do intervalo declarado deletado e a chamada decorre do alinhamento de um *contig* vizinho, que representa as regiões flanqueadoras, mas não o bloco gênico. A Tabela 1 resume o efeito das etapas sucessivas de filtragem.

Tabela 1 - Efeito dos filtros sucessivos sobre o conjunto bruto de SVs.

Etapa	Critério	Remanescentes
Chamada bruta (SVIM-asm)	SV ≥ 50 pb	6.298 SVs (7.177 registros)
Filtro de borda	> 2 kb das extremidades do *contig*	1.786 registros
Confirmação por leituras	SV concordante nas leituras (Sniffles2)	1.599 registros
Interseção com genes	SV sobrepondo modelo gênico	147 genes (156 pares)
Conjunto de maior confiança	Núcleo do *contig* + nº de leituras de suporte	76 genes (80 pares)
Sobreposição com região codificadora	SV incidindo sobre CDS	16 genes
Verificação manual	Exclusão de SVs que contêm o gene integralmente	8 genes

Fonte: elaborado pelo autor.

O conjunto de maior confiança, definido pela posição da variante no núcleo do *contig* e pelo maior número de leituras de suporte, reúne 80 pares em 76 genes e concentra os candidatos passíveis de interpretação. Dentro dele, 16 genes apresentaram variante sobreposta a região codificadora. Seis correspondem a deleções que englobam o gene por completo, com o gene presente na montagem (cobertura de 0,954 a 1,0), e dois consistem em inversões cujos pontos de quebra situam-se fora do gene, de modo que a sequência codificadora permanece contínua. Nos oito casos, a SV contém o gene sem interrompê-lo, sendo descartados da lista de SVs candidatas. Os oito genes restantes, nos quais a SV possui ponto de quebra interno à sequência codificadora, estão listados na Tabela 2.

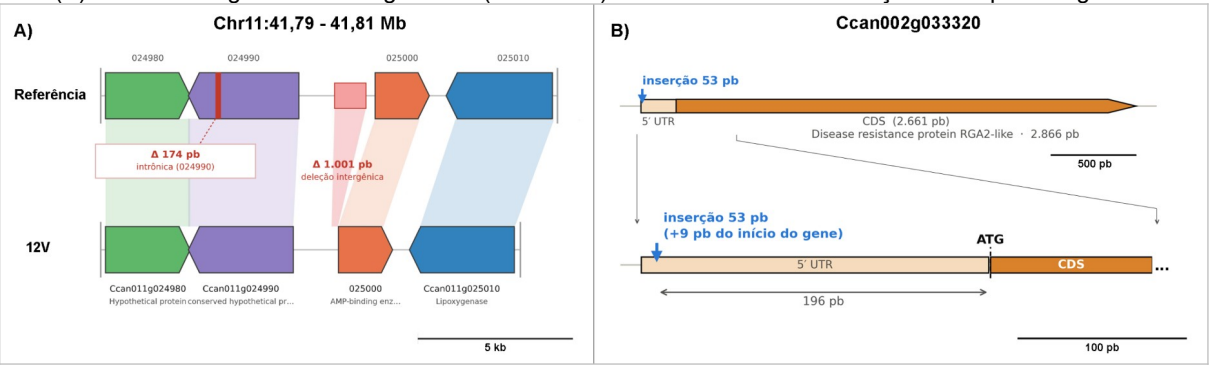
Tabela 2 - SVs localizadas na CDS de genes do clone 12V de *C. canephora*.

Gene	Tipo de SV	Tamanho do gene	Tamanho da SV	Função putativa
Ccan002g033280	INS	4.301 pb	682 pb	Protein DETOXIFICATION (MATE)
Ccan009g016950	DEL	2.385 pb	603 pb	DNA helicase dependente de ATP
Ccan002g026850	INS	2.189 pb	462 pb	PPR/PRORP
Ccan011g020250	INS	6.435 pb	247 pb	PPR At1g68930
Ccan006g027920	INS	730 pb	221 pb	Nuclease HARBI1
Ccan007g022340	DEL	6.679 pb	85 pb	Ankyrin repeat protein ITN1-like
Ccan010g007250	DEL	1.521 pb	70 pb	Peroxidase
Ccan008g021270	DEL	2.454 pb	63 pb	Proteína hipotética conservada

Fonte: elaborado pelo autor.

As demais SVs deste conjunto de maior confiança situam-se fora da sequência codificadora, em íntrons, regiões não traduzidas e segmentos intergênicos. Na região Chr11:41,79–41,81 Mb, a ordem e o sentido de quatro genes contíguos são mantidos, com duas deleções confirmadas por leituras: uma de 174 pb em íntron de Ccan011g024990 e uma de 1.001 pb na região intergênica entre Ccan011g024990 e Ccan011g025000 (Figura 1A). Destaca-se ainda uma inserção de 53 pb em Ccan002g033320, anotado como *Disease resistance protein RGA2-like*, localizada na porção 5′-UTR do gene, 196 pb a montante do códon de iniciação (Figura 1B).

Figura 1 - Loci representativos da variação estrutural no clone 12V de *C. canephora*. (A) Micro-colinearidade entre o clone 12V (*contig_7777*) e o genoma de referência na região Chr11:41,79–41,81 Mb; setas indicam o sentido dos genes, faixas ligam genes homólogos e blocos vermelhos marcam os segmentos ausentes no clone 12V. (B) Estrutura do gene Ccan002g033320 (*RGA2-like*) do clone 12V com a inserção de 53 pb na região 5′-UTR.



Fonte: elaborado pelo autor.

Discussão

Cerca de 75% das SVs identificadas na chamada bruta situavam-se a menos de 2 kb das extremidades dos *contigs*. Essas variantes foram eliminadas por filtragem pois, em uma montagem fragmentada, cada extremidade é um ponto em que o alinhamento termina abruptamente, de modo que o algoritmo de chamada tende a interpretar essa descontinuidade como variante estrutural, superestimando a ocorrência de SVs nesses locais (Meleshko et al., 2025; Qin; Li, 2025). Este filtro, contudo, não foi suficiente, uma vez que falsos positivos foram identificados ao confrontar a chamada com a anotação transferida pelo Liftoff, evidenciando a relevância desta etapa de verificação.

O tratamento das regiões repetitivas seguiu a lógica de que a resolução de regiões repetitivas é justamente a vantagem das *long-reads* (Smolka et al., 2024), razão pela qual genes derivados de transposons não foram descartados por princípio. Nestes casos, foram retidas as SVs sustentadas por leituras que atravessam os elementos, e descartadas as deleções de dezenas de quilobases sem nenhuma *long-read* de suporte, que têm maior possibilidade de serem artefatos de montagem. A permanência do gene Ccan006g027920, anotado como nuclease do tipo HARBI1, entre os oito genes finais ilustra a ambivalência dessa decisão. Trata-se de uma proteína derivada de um transposon da superfamília PIF/Harbinger. O modelo gênico possui 730 pb, não apresenta ORF válida, transfere-se com identidade de 0,716 e integra uma família com 84 cópias anotadas na referência. Em uma sequência com dezenas de cópias quase idênticas, a diferença observada pode refletir qual cópia foi amostrada, e não uma variante no locus. Assim, os elementos transponíveis são, simultaneamente, uma fonte de variação estrutural genuína, mas também o contexto no qual se concentram os artefatos, sendo que a distinção entre os dois casos não pode ser resolvida nesta cobertura.

Entre os oito genes com variante incidente sobre a sequência codificadora, a inserção de 682 pb em Ccan002g033280 (*DETOXIFICATION/MATE*) é a mais extensa, mas também a de menor sustentação, correspondendo a uma expansão de microssatélite $(GGAAT)_n$, e com baixa identidade de transferência da anotação pelo Liftoff (0,698). Regiões de baixa complexidade concentram desproporcionalmente os erros de chamada de variantes estruturais. No genoma humano, por exemplo, embora representem 1,2% da sequência, essas regiões abrigam mais da metade das chamadas obtidas por *long-reads* e de 77% a 91% das chamadas errôneas, atribuídas sobretudo a alinhamento inconsistente das leituras (Qin; Li, 2025). Esse caso deve, portanto, ser tratado como incerto e validado por outros métodos apesar da confirmação por leituras. Por outro lado, as duas proteínas PPR/PRORP,

envolvidas em processamento e edição de RNA em organelas, e a peroxidase apresentam sustentação mais favorável, com SVs de pequeno tamanho e internas a genes de estrutura preservada.

Fora do conjunto codificante, um caso de destaque entre os 76 genes de maior confiança consiste em uma inserção de 53 pb na porção 5′-UTR de Ccan002g033320, anotado como *Disease resistance protein RGA2-like* do tipo NBS-LRR, situada 196 pb a montante do códon de iniciação (Figura 1B). Ocorrem ainda duas proteínas quinases nesse conjunto, Ccan007g006980, do tipo BLUS1, e Ccan007g010950. Tais observações são compatíveis com a expectativa de que famílias gênicas de resistência e sinalização concentrem variação estrutural (Zhou et al., 2017; Liu et al., 2020), mas uma cobertura de 5× e uma montagem que captura menos de um terço do conteúdo gênico não permitem testar essa hipótese.

O conjunto identificado neste trabalho corresponde apenas à fração do genoma coberta pela montagem, e a perda de sensibilidade dos softwares na faixa de cobertura de 5× (Meleshko et al., 2025) implica que a ausência de uma variante em determinado locus não indica sua ausência no genoma. Desse modo, cada variante detectada deve ser considerada uma candidata, exigindo validação experimental.

Conclusão

Dados preliminares de uma única corrida de sequenciamento por nanoporos, ancorados ao genoma de referência de *C. canephora*, foram suficientes para localizar SVs entre um acesso de Conilon (clone 12V) e a referência. Foram identificados 156 pares gene-SV em 147 genes, dos quais oito apresentam ponto de quebra interno à sequência codificadora, além de uma inserção de 53 pb na região 5′-UTR de um gene de resistência do tipo NBS-LRR. Os genes candidatos podem ser validados por PCR, bem como por meio de outras corridas de sequenciamento para alcançar coberturas de 10-20×, permitindo estender a detecção de SVs para todo o genoma.

Referências

BAQUETA, M. R. et al. An overview on the Brazilian *Coffea canephora* scenario and the current chemometrics-based spectroscopic research. **Food Research International**, v. 194, p. 114866, 2024.

CABANETTES, F.; KLOPP, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. **PeerJ**, v. 6, p. e4958, 2018.

CUBRY, P. et al. Global analysis of *Coffea canephora* Pierre ex Froehner (Rubiaceae) from the Guineo-Congolese region reveals impacts from climatic refuges and migration effects. **Genetic Resources and Crop Evolution** v. 60, n. 2, p. 483-501, 2013.

DANECEK, P. et al. Twelve years of SAMtools and BCFtools. **GigaScience**, v. 10, n. 2, 2021.

DE COSTER, W. et al. NanoPack: visualizing and processing long-read sequencing data. **Bioinformatics**, v. 34, n. 15, p. 2666-2669, 2018.

DE OLIVEIRA, R. G. et al. Genomic discrimination of the botanical groups conilon and robusta of *Coffea canephora*. **Scientific Reports**, v. 16, n. 5584, 2026.

DENOEUD, F. et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. **Science**, v. 345, n. 6201, p. 1181-1184, 2014.

DEREEPER, A. et al. The coffee genome hub: a resource for coffee genomes. **Nucleic Acids Research**, v. 43, p. 1028-1035, 2015.

DOYLE, J. J.; DOYLE, J. L. Isolation of plant DNA from fresh tissue. **Focus**, v. 12, n. 1, p. 13-15, 1990.

FERRÃO, R. G. et al. **Café Conilon**. 2. ed. Vitória: Incaper, 2017. 784 p.

GROVER, C. E.; WENDEL, J. F. Recent insights into mechanisms of genome size change in plants. **Journal of Botany**, v. 2010, p. 1-8, 2010.

HELLER, D.; VINGRON, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. **Bioinformatics**, v. 36, n. 22-23, p. 5519-5521, 2020.

JAWO, T. O.; KYEREH, D.; LOJKA, B. The impact of climate change on coffee production of small farmers and their adaptation strategies: a review. **Climate and Development**, v. 15, n. 2, p. 93-109, 2023.

KOLMOGOROV, M. et al. Assembly of long, error-prone reads using repeat graphs. **Nature Biotechnology**, v. 37, n. 5, p. 540-546, 2019.

LI, H. Minimap2: pairwise alignment for nucleotide sequences. **Bioinformatics**, v. 34, n. 18, p. 3094-3100, 2018.

LIU, Y. et al. Pan-genome of wild and cultivated soybeans. **Cell**, v. 182, n. 1, p. 162-176, 2020.

MANNI, M. et al. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage. **Molecular Biology and Evolution**, v. 38, n. 10, p. 4647-4654, 2021.

MELESHKO, D. et al. Blackbird: structural variant detection using synthetic and low-coverage long-reads. **Bioinformatics Advances**, v. 5, n. 1, 2025.

PEDERSEN, B. S.; QUINLAN, A. R. Mosdepth: quick coverage calculation for genomes and exomes. **Bioinformatics**, v. 34, n. 5, p. 867-868, 2018.

POLKHOVSKAYA, E. et al. Cost-effective detection of SNPs and structural variations in full-length genes of wheat and sunflower using multiplex PCR and rapid Nanopore kit. **Biology**, v. 14, n. 2, p. 138, 2025.

QIN, Q.; LI, H. Challenges in structural variant calling in low-complexity regions. **GigaScience**, v. 14, n. 154, 2025.

QUINLAN, A. R.; HALL, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics**, v. 26, n. 6, p. 841-842, 2010.

SALOJÄRVI, J. et al. The genome and population genomics of allopolyploid *Coffea arabica* reveal the diversification history of modern coffee cultivars. **Nature Genetics**, v. 56, n. 4, p. 721-731, 2024.

SHUMATE, A.; SALZBERG, S. L. Liftoff: accurate mapping of gene annotations. **Bioinformatics**, v. 37, n. 12, p. 1639-1643, 2021.

SMOLKA, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. **Nature Biotechnology**, v. 42, n. 10, p. 1571-1580, 2024.

YUAN, Y. et al. Current status of structural variation studies in plants. **Plant Biotechnology Journal**, v. 19, n. 11, p. 2153-2163, 2021.

ZAIDAN, I. R. et al. Diversity and structure of *Coffea canephora* from old seminal crops in Espírito Santo, Brazil: genetic resources for coffee breeding. **Tree Genetics & Genomes**, v. 19, n. 2, 2023.

ZHOU, P. et al. Exploring structural variation and gene family architecture with de novo assemblies of 15 *Medicago* genomes. **BMC Genomics**, v. 18, n. 1, 2017.

Agradecimentos

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (Fapes) – 378/2026 -P: 2026-1XCKD. Os autores também agradecem à Universidade Federal do Espírito Santo (UFES), ao Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (Incaper), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio institucional.