



OPEN ACCESS

EDITED BY

Carla M. R. Varanda,
Research Centre for Natural Resources,
Environment and Society (CERNAS),
Portugal

REVIEWED BY

Huawei Tan,
Nanjing University, China
Xuanmin Guang,
Beijing Genomics Institute (BGI), China

*CORRESPONDENCE

Ezekiel Ahn
ezekiel.ahn@usda.gov

RECEIVED 02 July 2026
REVISED 09 August 2026
ACCEPTED 17 August 2026
PUBLISHED 31 August 2026

CITATION

Ahn E, Park S, Baek I, Kim MS and
Meinhardt LW (2026) Genome-wide
subgenome-resolved analysis validates
chromosome 4 differentiation and
prioritizes introgressed
Coffea arabica accessions.
Front. Plant Sci. 17:1925916.
doi: 10.3389/fpls.2026.1925916

COPYRIGHT

© 2026 Ahn, Park, Baek, Kim and
Meinhardt. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication
in this journal is cited, in accordance
with accepted academic practice. No
use, distribution or reproduction is
permitted which does not comply with
these terms.

Genome-wide subgenome-resolved analysis validates chromosome 4 differentiation and prioritizes introgressed *Coffea arabica* accessions

Ezekiel Ahn^{1*}, Sunchung Park¹, Insuck Baek², Moon S. Kim¹
and Lyndel W. Meinhardt¹

¹Sustainable Perennial Crops Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD, United States, ²Environmental Microbial and Food Safety Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD, United States

Chromosome 4 introgression in Timor hybrid-derived *Coffea arabica* is established, but the robustness of accession prioritization and the relative strength of cultivated–introgressed differentiation across the canephora-derived (sgC) and eugenioides-derived (sgE) subgenomes remained unclear under explicit subgenome filtering. We reanalyzed public genomic resources from 44 coffee accessions using strict contig-level subgenome filtering, Arabica-only population-structure analysis, SNP-panel sensitivity testing, genome-wide differentiation scans, permutation testing, and direct sequence alignment. Population structure and accession rankings were stable across marker densities and random seeds, and the same six introgressed references were retained throughout. Chromosome 4 ranked first in both subgenomes, with a strong sgC signal and a markedly weaker sgE signal; independent baseline-panel permutation tests supported both chromosome 4-associated signals. Direct alignment supported correspondence to the expected chromosome 4 pseudomolecules while showing incomplete source coverage and unresolved exact boundaries. Alignment-supported blocks contained 88 sgC and 62 sgE provisional defense-, signaling-, and regulatory-associated annotations. These results provide a genome-wide, quantitatively validated framework for prioritizing introgressed germplasm and candidate chromosome 4 regions for phenotype-linked coffee research without implying equivalent introgression, exact liftover, or causal resistance genes.

KEYWORDS

allopolyploidy, chromosome 4, *Coffea arabica*, introgression, population genomics, sensitivity analysis, subgenome filtering

1 Introduction

The genus *Coffea* includes diverse diploid species, whereas cultivated coffee production is dominated by diploid *C. canephora* and allotetraploid *C. arabica*. Arabica originated through hybridization between ancestors related to *C. canephora* and *C. eugenioides*, leaving modern accessions with a canephora-derived subgenome (sgC) and an eugenioides-derived subgenome (sgE) (Merot-L'anthoene et al., 2019; Scalabrin et al., 2020; Salojärvi et al., 2024).

Its recent origin, dispersal through a small number of cultivated lineages, and subsequent bottlenecks produced an unusually narrow genetic base (Anthony et al., 2002; Gaut et al., 2018), which constrains improvement under changing climates and persistent disease pressure (Ngure and Watanabe, 2024).

Introgression has therefore been central to Arabica improvement. Hibrido de Timor and related materials, derived from *C. arabica*–*C. canephora* hybridization, have contributed disease-resistance-associated diversity to Catimor-, Sarchimor-, and other breeding backgrounds (Lashermes et al., 2000; Herrera et al., 2002; Clarindo et al., 2013; Van der Vossen et al., 2015). Such germplasm can also carry quality or agronomic trade-offs, making accession identity and genomic context important for downstream breeding decisions (Bertrand et al., 2003; Lashermes et al., 2011). More broadly, studies of perennial domestication and crop-wild gene flow show that introgression can contribute adaptive variation while leaving strongly nonuniform genomic signatures (Hufford et al., 2013; Gaut et al., 2015).

Salojärvi et al. (2024) established a shared Timor-derived chromosome 4 introgression region, primarily in sgC, and connected that region with resistance-associated gene space. Accurate subgenome assignment is essential in allopolyploids because homoeologous exchange, structural variation, and analytical misassignment can distort subgenome-specific evolutionary inference (Deb et al., 2023; Scalabrin et al., 2024; Session, 2024). The present study does not claim *de novo* discovery of that signal. Instead, it addresses four complementary questions that were not quantified in the source analysis: whether chromosome 4 remains the strongest signal after strict contig-level subgenome filtering; whether the result is stable across SNP panel sizes and random seeds; which introgressed accessions are prioritized reproducibly; and whether provisional annotation context is supported by direct sequence alignment rather than proportional coordinate scaling alone. The weaker sgE chromosome 4 signal reported below is described as matched group differentiation because it is observed in the same cultivated-introgressed accession contrast; this descriptive term does not imply physical linkage, equivalent alien introgression, or a shared causal mechanism.

Table 1 contrasts the previously reported biological discovery with the quantitative validation framework and reusable outputs. The workflow explicitly separates sgC- and sgE-labelled pseudomolecules, uses Arabica-only PCA for within-Arabica inference, evaluates sampling and reference-panel sensitivity, performs

an unbiased genome-wide scan, and restricts candidate annotations to direct alignment-supported blocks.

Accordingly, we reanalyzed the public 44-accession panel to (i) quantify strict subgenome-specific population structure, (ii) test the stability of chromosome-level differentiation and accession ranking, (iii) compare full-panel and Arabica-only ordinations, and (iv) replace proportional annotation projection with direct sequence-alignment-supported annotation summaries.

2 Materials and methods

2.1 Public genomic resources and study design

The analysis used public resources associated with Salojärvi et al. (2024): accession metadata and two Arabica variant-call files distributed as sgC and sgE callsets. The panel comprised 39 *C. arabica* accessions classified as cultivated, introgressed, or wild, together with three *C. canephora* and two *C. eugenioides* reference accessions. ET-39 sgC and sgE genome assemblies and GFF3 annotations were obtained from the public *Coffea* assembly collection, and assembly metadata were retrieved through NCBI Datasets. No new plant material, sequencing, or phenotype data were generated.

2.2 Metadata harmonization, accession context, and strict subgenome filtering

Accession identifiers were reconciled by case- and punctuation-insensitive normalization and conservative removal of source-specific suffixes. Distributed metadata supplied accession name, country or source, donor institute, pedigree notes, and varietal classification. Biological context for the six prioritized introgressed references was compiled from those metadata and relevant coffee-breeding literature; it is presented as contextual information only because no disease, quality, or agronomic phenotypes were measured in this study (Supplementary Table S1).

Both public VCFs were called against a combined Arabica reference and contained sgC-labelled pseudomolecules, sgE-labelled pseudomolecules, and unassigned scaffolds. The workflow therefore restricted the sgC analysis to the 11 contigs explicitly labelled “_sg_C_”

TABLE 1 Relationship between the previously reported chromosome 4 introgression result and the quantitative validation framework.

Component	Previously reported study (Salojärvi et al., 2024)	Quantitative validation framework
Chromosome 4 result	Timor-derived <i>C. canephora</i> introgression was identified primarily in sgC.	An unbiased genome-wide scan quantitatively validates chromosome 4 as the top pseudomolecule after strict subgenome filtering.
Subgenome interpretation	Direct introgression was interpreted primarily in sgC.	sgC retains a strong signal; sgE shows a weaker chromosome 4-associated differentiation signal in the same accession contrast, without a claim of equivalent introgression.
Robustness analysis	Population-genomic and expression evidence established the biological region.	Four SNP panel sizes, five seeds, all-eligible PCA, Arabica-only sensitivity, all-eligible genome-wide scanning, and 500 label permutations of the strict-filter 12,000-SNP baseline quantify robustness.
Accession-level output	Timor-derived materials were classified as introgressed or hybrid accessions.	A stable top-six accession ranking is recovered in every sensitivity condition, with geographic and pedigree context summarized in Supplementary Table S1.
Coordinate and annotation output	Resistance-associated gene space was reported in the chromosome 4 region.	Direct sequence alignment validates chromosome-level correspondence and defines alignment-supported provisional candidate annotations; exact boundaries and causality are not inferred.

and the sgE analysis to the 11 contigs explicitly labelled “_sg_E_” before any PCA, ranking, or differentiation analysis. Opposite-subgenome pseudomolecules and unassigned scaffolds were excluded. The audit retained 23,225,336 sgC-labelled and 25,062,481 sgE-labelled records and excluded 5,768,784 and 6,991,157 opposite-subgenome records, respectively (Supplementary Table S2).

2.3 Variant filtering and SNP-panel sensitivity

Analyses were restricted to biallelic SNPs with site call rate at least 0.80 and minor allele frequency at least 0.05, yielding 3,601,265 eligible sgC variants and 5,317,075 eligible sgE variants. A strict-filter 12,000-SNP reservoir sample generated with seed 20250416 served as an equal-density operational baseline for repeated PCA, accession-ranking, interval, and permutation analyses (Vitter, 1985). Using a fixed marker count provided matched marker density between sgC and sgE despite their different numbers of eligible variants and made repeated sensitivity and permutation calculations directly comparable. The baseline was not used as a substitute for the complete filtered datasets: all eligible variants were included in the streaming PCA and genome-wide differentiation scans. Its adequacy was evaluated against deterministic bottom-k samples of 6,000, 12,000, 24,000, and 48,000 SNPs under five independent seeds (20250416–20250420). Stability was evaluated by pairwise-distance correlations, Procrustes similarity, interval overlap, accession-rank correlations, and top-six membership (Supplementary Tables S3, S4; Supplementary Figure S1).

2.4 Population structure and diploid-reference sensitivity

PCA was performed on standardized diploid dosage matrices using five components (Jolliffe and Cadima, 2016). Because the five diploid references are deeply divergent from Arabica and numerically unbalanced, primary inference on within-Arabica structure used the 39-accession Arabica-only panel. The full 44-accession analysis was retained as progenitor context. For each sensitivity condition, Arabica pairwise distances in the two analyses were correlated, and cultivated-introgressed centroid separation was compared (Supplementary Table S5; Supplementary Figure S2).

2.5 Genome-wide differentiation and permutation testing

For the cultivated-versus-introgressed contrast, absolute allele-frequency difference ($|\Delta AF|$) was calculated at every eligible SNP. All-eligible genome-wide summaries were generated for each labelled pseudomolecule and for non-overlapping 1-Mb windows containing at least five valid markers. Chromosomes were ranked by mean $|\Delta AF|$, while operational interval summaries used the locations of the 250 most differentiated markers. Separately, significance was evaluated at the 12,000-SNP baseline marker density by permuting cultivated and introgressed labels 500 times within the strict-filter 12,000-SNP baseline panel while preserving group sizes. The maximum 1-Mb window mean $|\Delta AF|$ was recorded for each permutation. Empirical P values were calculated as $(b + 1)/$

$(500 + 1)$, where b is the number of permuted maxima at least as large as the observed baseline-panel maximum (Supplementary Table S6; Supplementary Figure S3).

2.6 Accession prioritization

Accession-level introgression affinity was calculated in the first five Arabica-only principal components. For each accession, Euclidean distance to the cultivated centroid minus distance to the introgressed centroid was used as the affinity score; larger values therefore indicate a position closer to the introgressed than to the cultivated centroid. Introgressed accessions were ranked within each subgenome and sensitivity condition. Robustness was summarized by Spearman rank correlation with the filtered 12,000-SNP baseline and by overlap of the six highest-ranked introgressed accessions (Supplementary Table S7).

2.7 Direct interval alignment and provisional annotation context

Operational chromosome 4 source intervals from the baseline analysis were extracted from assembly GCA_036785775.1 and divided into 200-kb chunks with 20-kb overlap. Chunks were aligned by NCBI BLAST+ megablast to the corresponding ET-39 sgC and sgE assemblies, retaining high-scoring segment pairs at least 1 kb long and at least 85% identical (Camacho et al., 2009). For each subgenome we quantified source coverage, weighted identity, orientation, source-target order correlation, and the largest alternative-chromosome coverage. This analysis evaluated chromosome-level correspondence and local collinearity; it was not treated as a single exact base-pair liftover.

GFF3 genes overlapping direct alignment-supported target blocks were extracted as the supported gene inventory. Provisional candidates were defined by annotation descriptors associated with defense receptors, kinase signaling, stress response, transcriptional regulation, membrane transport, or leucine-rich repeats. Candidate annotations were not interpreted as causal loci, and exact interval endpoints were not inferred (Supplementary Tables S8–S10).

2.8 Computational environment and reproducibility

Analyses were conducted in Python using NumPy for numerical operations, pandas for tabular processing, and openpyxl for workbook export (McKinney, 2010; Harris et al., 2020). The public repository contains the strict subgenome-filtering, sensitivity, genome-wide differentiation, accession-ranking, direct-alignment, and annotation-export workflows, together with machine-readable source data underlying all figures and supplementary tables.

3 Results

3.1 Arabica-only population structure after strict subgenome filtering

After explicit contig-level filtering, both sgC and sgE Arabica-only PCA separated the principal introgressed references from the

cultivated and wild accessions (Figure 1). Inclusion of the three *C. canephora* and two *C. eugenoides* references changed detailed within-Arabica ordination geometry: the median correlation between full-panel and Arabica-only Arabica distances was 0.860 across sensitivity conditions. However, cultivated-introgressed centroid separation increased in the Arabica-only analysis by approximately 31% in sgC and 83% in sgE. We therefore use Arabica-only PCA for within-Arabica inference and retain the full-panel ordinations only as supplementary progenitor context (Supplementary Table S5; Supplementary Figure S2).

3.2 All-eligible scans support a strong sgC and weaker sgE chromosome 4 signal

The all-eligible analysis was genome-wide rather than chromosome 4-targeted. Chromosome 4 was the top-ranked pseudomolecule in both strict subgenome scans (Figure 2; Supplementary Table S6). In sgC, chromosome 4 had a mean $|\Delta AF|$ of 0.3878 and contained 46,552 markers with $|\Delta AF|$ at least 0.90; the all-eligible maximum-mean 1-Mb window was 14–15 Mb (mean $|\Delta AF| = 0.5450$). In sgE, chromosome 4 also ranked first but the signal was markedly weaker: mean $|\Delta AF|$ was 0.0843, 642 markers had $|\Delta AF|$ at least 0.90, and the all-eligible maximum-mean window was 11–12 Mb (mean $|\Delta AF| = 0.2137$). In separate 500-permutation tests of the strict-filter 12,000-SNP baseline, the observed maximum window means were 0.5738 at 14–15 Mb in sgC and 0.2232 at 17–18 Mb in sgE, exceeding null 99th percentiles of 0.3482 and 0.1656, respectively (empirical $P = 0.001996$ and 0.003992). Thus, both subgenomes show chromosome 4-associated differentiation in the same accession contrast, but the magnitude is strongly asymmetric and does not support equivalent introgression.

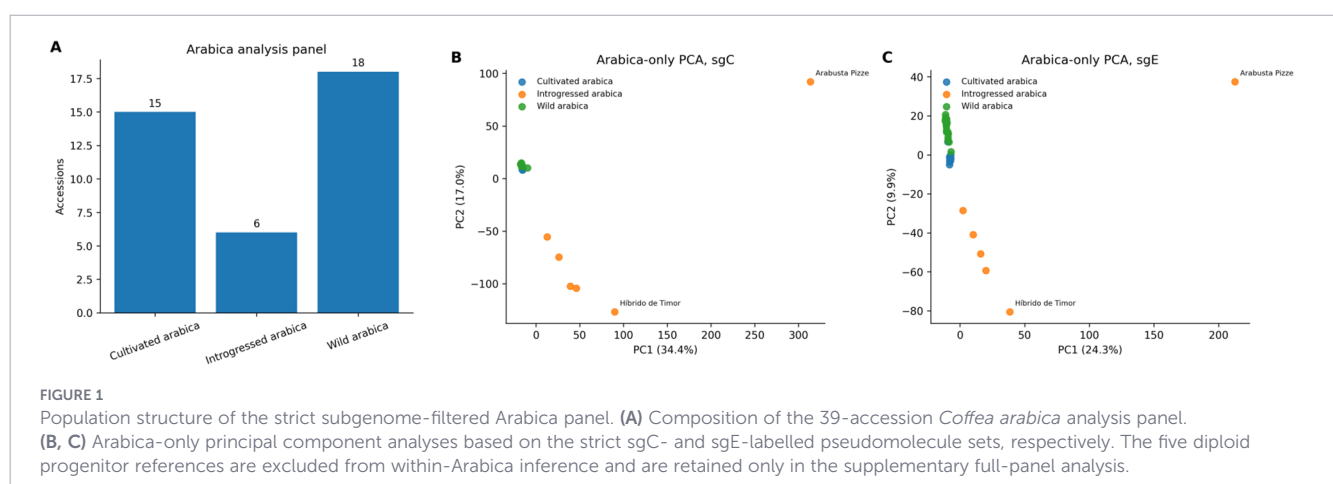
3.3 SNP-panel sensitivity and accession prioritization are reproducible

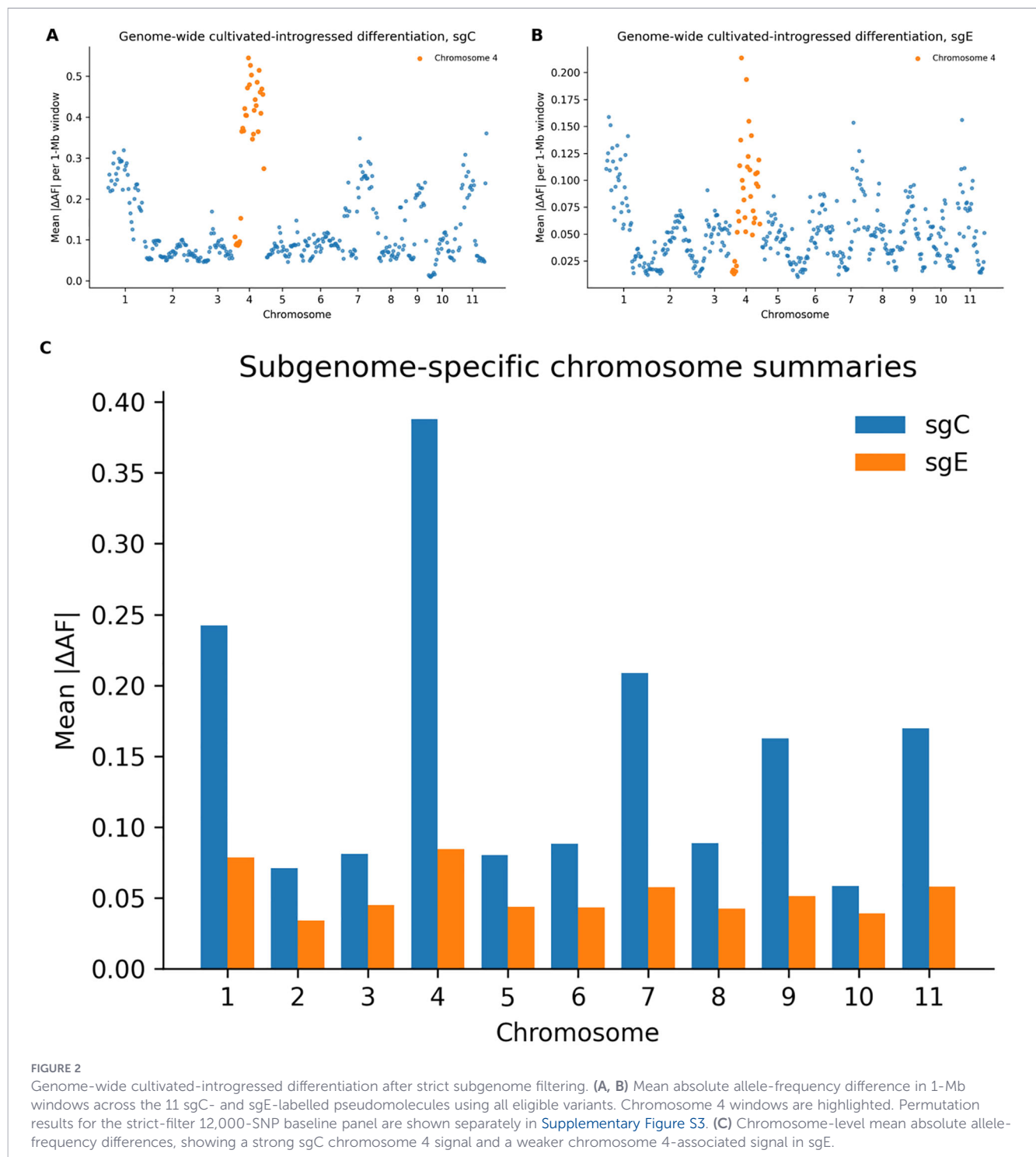
PCA structure was stable across panel sizes, seeds, and the all-eligible analysis (Supplementary Table S4; Supplementary Figure S1). The 12,000-SNP density was the smallest tested condition that recovered chromosome 4 as the top-ranked pseudomolecule in all five seeds for both subgenomes. The 6,000-SNP condition was less

stable, particularly in sgE, whereas increasing the density to 24,000 or 48,000 SNPs did not materially alter population structure, chromosome ranking, or accession prioritization. Results from the complete filtered variant sets were likewise concordant with the 12,000-SNP baseline: Arabica-only PCA pairwise-distance correlations were 0.9955 for sgC and 0.9996 for sgE, and the same six introgressed accessions remained in the top six under all tested conditions. We therefore retained 12,000 SNPs as the smallest empirically sufficient equal-density baseline rather than as a claim of universal optimality.

Accession prioritization was likewise stable. Median Spearman correlation of accession affinity with the filtered 12,000-SNP baseline was 0.9729, and the same six introgressed references were retained in every condition: Híbrido de Timor, Arabusta Pizze, Iapar 59, Oro Azteca, IPR99, and Costa Rica 95 (Figure 3; Supplementary Table S7). These accessions include a natural Timor-derived donor, an inter-specific Arabusta background, and Catimor- or Sarchimor-derived cultivars distributed through Brazilian and French donor collections. Their geographic source, pedigree context, and published or metadata-based disease and agronomic information are summarized in Supplementary Table S1. For example, IPR99 has been reported as resistant to coffee ringspot virus and partially resistant to leaf rust, but no phenotype was measured for any accession in the present study (Sera et al., 2013).

The subgenome-specific affinity profiles further distinguish the principal reference backgrounds. Híbrido de Timor ranked first in sgC under all 22 evaluated conditions (baseline affinity score, 70.66), whereas Arabusta Pizze ranked first in sgE (44.43) and second in sgC. Iapar 59 ranked third in sgC and alternated between second and third in sgE; Oro Azteca, IPR99, and Costa Rica 95 remained fourth, fifth, and sixth, respectively, in both subgenomes. These results identify Híbrido de Timor and Arabusta Pizze as the strongest accession-level anchors of the sgC and sgE components of the cultivated–introgressed contrast, while the four derived cultivars provide stable downstream reference points. The chromosome 4 regions identified by the independent genome-wide scans provide a concrete genomic focus for comparative follow-up among these accessions; however, the present analysis does not infer phased accession-specific introgression boundaries, and the affinity scores do not measure field resistance or agronomic performance.





3.4 Direct alignment supports chromosome-level correspondence and a provisional annotation resource

Direct sequence alignment assigned all highest-scoring source chunks to the expected ET-39 chromosome 4 pseudomolecules. Retained alignments covered 70.29% of the sgC source interval and 70.39% of the sgE source interval, with weighted identities of 99.643% and 99.653%, respectively (Figure 4; [Supplementary Table S10](#)). Source–target order correlations were 0.9993 in sgC

and 0.9999 in sgE, while the largest alternative-chromosome coverage was below 4% in both analyses. These results support chromosome-level correspondence and argue against wholesale relocation to another chromosome. Incomplete coverage and a minority of reverse-orientation blocks nevertheless preclude interpretation as single base-pair-resolved liftover intervals.

Alignment-supported blocks contained 900 sgC and 692 sgE gene models, of which 88 and 62, respectively, met the provisional candidate-annotation criteria ([Supplementary Tables S8, S9](#)). The retained annotations included RPP13-like and RPP8-like resistance

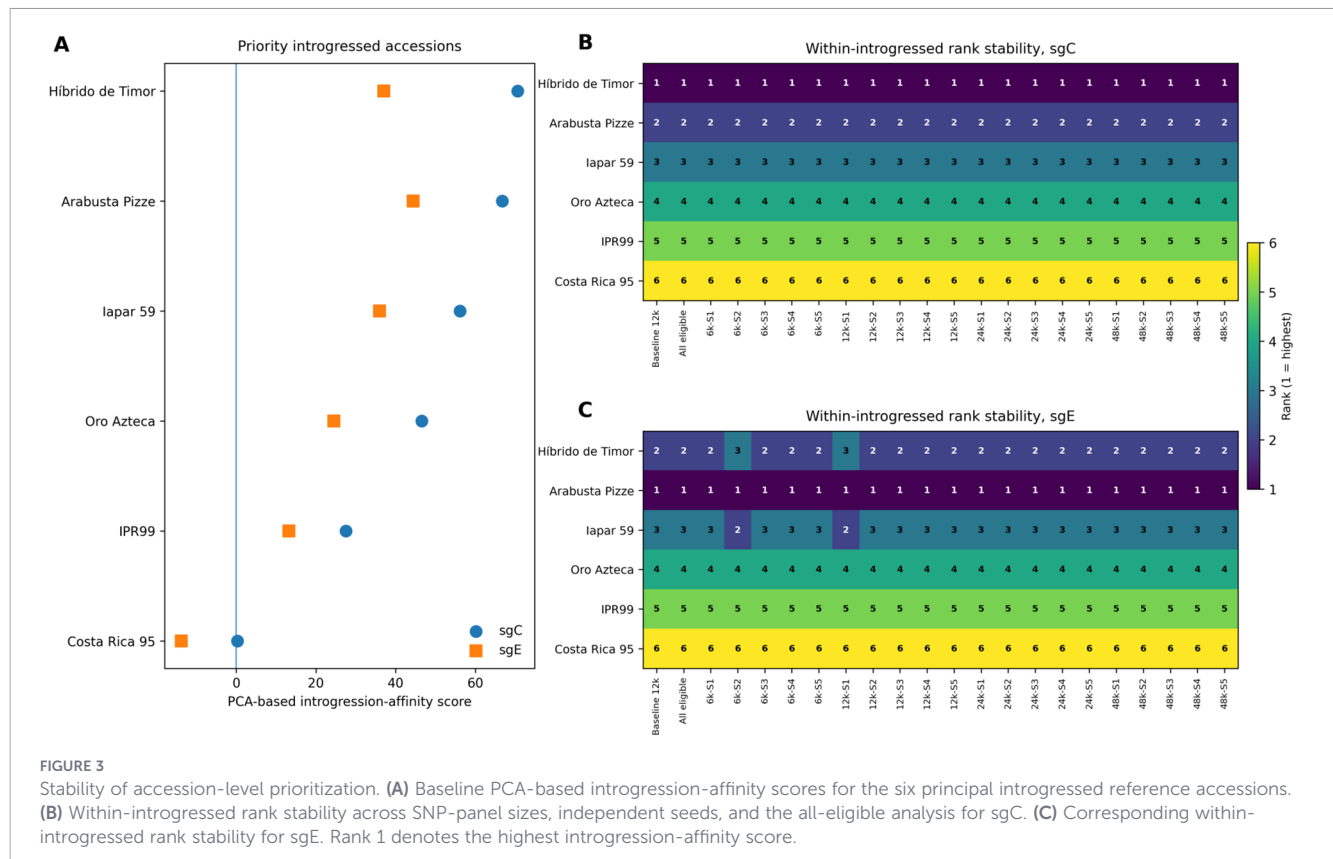


FIGURE 3

Stability of accession-level prioritization. (A) Baseline PCA-based introgression-affinity scores for the six principal introgressed reference accessions. (B) Within-introgressed rank stability across SNP-panel sizes, independent seeds, and the all-eligible analysis for sgC. (C) Corresponding within-introgressed rank stability for sgE. Rank 1 denotes the highest introgression-affinity score.

proteins, receptor-like kinase and WNK3-associated loci, JA2L-like and other NAC-domain annotations, and leucine-rich-repeat-associated proteins. These features provide biologically plausible annotation context, but they do not establish resistance phenotypes or causal gene function.

4 Discussion

This study does not rediscover the Timor-associated chromosome 4 introgression reported by Salojärvi et al. (2024). Its incremental contribution is a quantitative validation layer: strict contig-level subgenome filtering, all-eligible genome-wide scanning, independent permutation testing of the strict-filter 12,000-SNP baseline, multi-panel sampling sensitivity, Arabica-only ordination, stable accession ranking, and direct sequence-alignment-supported annotation. Together, these analyses convert a known biological signal into a reproducible resource whose strongest and weakest claims are explicitly separated.

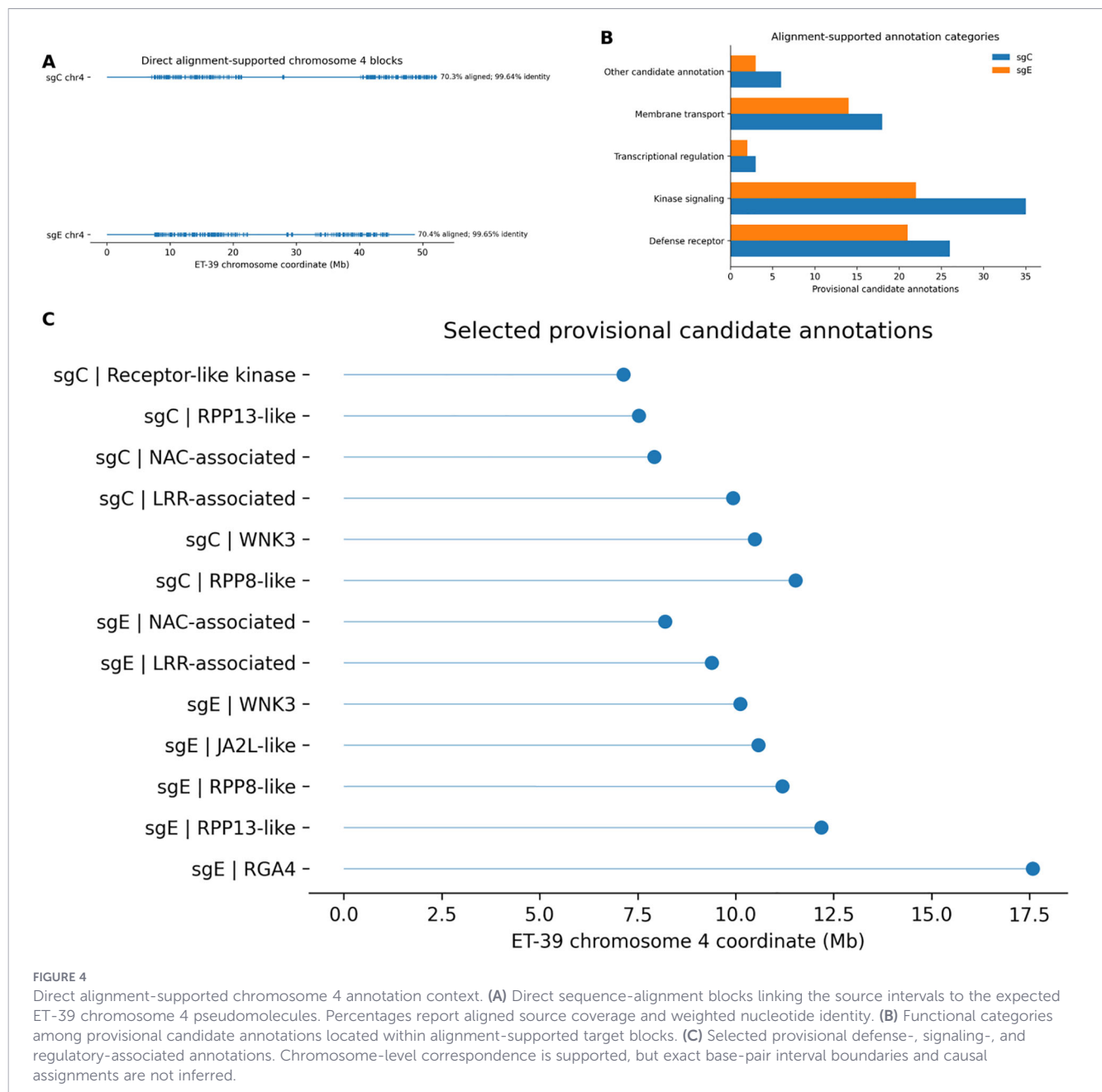
Explicit subgenome filtering was essential because both public VCFs use a combined Arabica reference and contain sgC-labelled, sgE-labelled, and unassigned contigs. Treating file identity alone as subgenome identity can mix records from the two genomic compartments. The workflow therefore excludes opposite-subgenome and unassigned records before all downstream analyses.

The strict all-eligible genome-wide scan strongly supports the expected sgC result. Chromosome 4 ranked first genome-wide and was recovered in every 12,000-SNP-or-larger replicate; the independent strict-filter 12,000-SNP baseline permutation test also

placed its observed maximum above the null distribution. Subgenome-specific breeding signatures have likewise been documented in other allopolyploid crops, where selection can preserve markedly different genomic patterns across homoeologous compartments (Qian et al., 2014; Wang et al., 2022). The sgE chromosome 4 signal also ranked first in the all-eligible scan and was stable at 12,000 SNPs or more, while its baseline-panel permutation maximum exceeded the corresponding null distribution. Its mean effect and number of extreme markers were nevertheless far smaller. We therefore interpret sgE as weaker chromosome 4-associated group differentiation observed in the same accession contrast, not as evidence of direct *C. canephora* introgression, physical linkage to sgC, or a shared causal block. This interpretation is consistent with the source study's conclusion that recent alien introgression occurred predominantly in sgC.

The accession-ranking output is one of the most reproducible additions. The same six introgressed references were retained across every panel size, seed, and all-eligible analysis. Their backgrounds also make biological sense: Híbrido de Timor anchors the historically important donor lineage; Arabusta Pizze represents a tetraploid Arabica-canephora hybrid background; and Iapar 59, Oro Azteca, IPR99, and Costa Rica 95 represent Timor-derived breeding lineages. Nevertheless, ranking here measures genomic position relative to the cultivated and introgressed centroids, not field performance. The analysis does not establish that a higher affinity score predicts rust resistance, quality, yield, or adaptation.

Direct alignment substantially improves the annotation analysis compared with proportional coordinate scaling. The high nucleotide identities, near-monotonic source-target order, expected



chromosome assignments, and low alternative-chromosome coverage support chromosome-level correspondence. At the same time, approximately 30% of each operational source interval lacked retained direct alignment coverage, and some blocks were reversed. Exact boundaries therefore remain unresolved, and the target-coordinate envelopes should not be interpreted as continuous orthologous interval sizes. Restricting the gene inventory to alignment-supported blocks yields a more defensible provisional annotation library, including resistance-protein, receptor-like kinase, NAC, and leucine-rich-repeat classes that are broadly compatible with plant immune signaling (Wu and Zhou, 2013; Stephens et al., 2022; Dong et al., 2024). Coffee genomes contain large NLR repertoires, and comparative resequencing has identified disease-resistance gene families among the most variable genomic compartments across *Coffea* species (Huang et al., 2020; Santos et al.,

2022). A recent study also functionally validated a rust race II resistance-associated region in *C. arabica*, illustrating the phenotype-linked follow-up required before candidates from the present analysis can be interpreted functionally (Ariyoshi et al., 2024).

The host-genomic context is relevant to coffee leaf rust, but the present study does not directly test host-pathogen interactions. *Hemileia vastatrix* populations show host adaptation, recombination, and geographic structure (Silva et al., 2018; Rodrigues et al., 2022), and Timor-derived breeding materials have repeatedly been used in rust-resistance improvement. However, no inoculation, transcriptomic, field, or gene-function data were generated here. The accession context and candidate annotations should therefore guide, not substitute for, phenotype-linked validation.

Several limitations remain. First, the study relies on processed public VCFs and inherits their upstream variant-calling choices.

Second, although sensitivity analysis demonstrates that the main population structure, chromosome ranking, and accession ranking are stable, operational interval endpoints vary with marker density and should not be interpreted as fixed fine-mapping boundaries. Third, Arabica ordination geometry is influenced by inclusion of the five deeply divergent diploid references; this is why the Arabica-only PCA is used for primary inference. Fourth, direct alignment provides approximately 70% source coverage rather than complete lift-over, leaving unresolved sequence and possible local structural differences. Finally, all gene and breeding interpretations remain hypothesis-generating until tested with phenotype, expression, or functional data.

5 Conclusion

Strict subgenome filtering, all-eligible genome-wide scans, and independent permutation tests of the strict-filter 12,000-SNP base-line confirm that chromosome 4 is the dominant cultivated-introgressed differentiation signal in sgC and a weaker but significant signal in sgE. The result is stable across SNP panel sizes and seeds, and the six principal introgressed reference accessions are retained under every sensitivity condition. Direct sequence alignment supports chromosome-level correspondence and an alignment-supported provisional annotation resource, but not exact interval boundaries or causal resistance genes. This analysis therefore links the strong sgC and weaker sgE chromosome 4 signals to six consistently prioritized introgressed references. Hibrido de Timor and Arabusta Pizze provide the strongest sgC and sgE affinity anchors, respectively, while Iapar 59, Oro Azteca, IPR99, and Costa Rica 95 provide stable derived breeding references for future chromosome 4 haplotype and phenotype comparisons. These rankings reflect genomic affinity rather than measured field resistance or agronomic performance.

Data availability statement

All datasets analyzed in this study are publicly available. Subgenome-resolved variant calls, accession metadata, syntenic alignments, and documentation are available from Dryad at <https://doi.org/10.5061/dryad.qnk98sfpt> (Salojärvi, 2023). ET-39 subgenome assemblies and GFF3 annotations are available from Zenodo at <https://doi.org/10.5281/zenodo.17250347> (Sotero and Domingues, 2025), and assembly metadata are available through NCBI Datasets under accession GCA_036785775.1 (National Center for Biotechnology Information, 2024). No new sequencing, genotype, phenotype, or experimental data were generated. The analysis workflow for strict subgenome filtering, sensitivity testing, genome-wide differentiation, accession prioritization, direct interval alignment, and annotation export is available at <https://github.com/EJSAHN/coffee-arabica-introgression-architecture>. Machine-readable source data underlying all figures and supplementary tables are provided with the repository and article.

Author contributions

EA: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review & editing. SP: Formal analysis, Methodology, Validation, Writing – review & editing. IB: Methodology, Validation, Writing – review & editing. MSK: Resources, Supervision, Writing – review & editing. LWM: Funding acquisition, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the U.S. Department of Agriculture, Agricultural Research Service, In-House Projects Nos. 8042-21220-258-000-D and 8042-21000-303-000-D.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer HT declared a past co-authorship with the author (s) SP and LM to the handling editor.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. ChatGPT and Claude were used to assist with code refactoring and language editing. The authors reviewed all content and take full responsibility for the final manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer, and all agency services are available without discrimination.

Supplementary material

Additional file 1: [Supplementary Tables S1–S10](#), including the accession context, strict contig-filter audit, validation overview, SNP/PCA sensitivity, diploid-reference sensitivity, separate all-eligible

genome-wide and strict-filter 12,000-SNP baseline permutation results, accession-rank stability, alignment-supported gene inventory, provisional candidate annotations, and direct-alignment diagnostics.

Additional file 2: [Supplementary Data S1](#), containing the expanded machine-readable outputs underlying the figures, supplementary tables, strict-filter audits, PCA analyses, accession rankings, operational interval summaries, all-eligible genome-wide scans, strict-filter 12,000-SNP baseline permutation tests, direct-alignment blocks, and annotation-supported gene lists.

Additional file 3: [Supplementary Figures S1–S3](#) showing SNP-subsampling stability, diploid-reference sensitivity, and permutation validation of the strict-filter 12,000-SNP baseline panel.

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2026.1925916/full#supplementary-material>

References

- Anthony, F., Combes, M., Astorga, C., Bertrand, B., Graziosi, G., and Lashermes, P. (2002). The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. *Theor. Appl. Genet.* 104, 894–900. doi: 10.1007/s00122-001-0798-8
- Ariyoshi, C., Ivamoto-Suzuki, S. T., Baba, V. Y., Toma-Braghini, M., Sera, G. H., Powell, A., et al. (2024). Characterization and functional validation of a genomic region involved in resistance to rust race II in *Coffea arabica*. *Australas. Plant Pathol.* 53, 363–373. doi: 10.1007/s13313-024-00988-2
- Bertrand, B., Guyot, B., Anthony, F., and Lashermes, P. (2003). Impact of the *Coffea canephora* gene introgression on beverage quality of *C. arabica*. *Theor. Appl. Genet.* 107, 387–394. doi: 10.1007/s00122-003-1203-6
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinform.* 10, 421. doi: 10.1186/1471-2105-10-421
- Clarindo, W. R., Carvalho, C. R., Caixeta, E. T., and Koehler, A. D. (2013). Following the track of “Híbrido de Timor” origin by cytogenetic and flow cytometry approaches. *Genet. Resour. Crop Evol.* 60, 2253–2259. doi: 10.1007/s10722-013-9990-3
- Deb, S. K., Edger, P. P., Pires, J. C., and McKain, M. R. (2023). Patterns, mechanisms, and consequences of homoeologous exchange in allopolyploid angiosperms: a genomic and epigenomic perspective. *New Phytol.* 238, 2284–2304. doi: 10.1111/nph.18927
- Dong, B., Liu, Y., Huang, G., Song, A., Chen, S., Jiang, J., et al. (2024). Plant NAC transcription factors in the battle against pathogens. *BMC Plant Biol.* 24, 958. doi: 10.1186/s12870-024-05636-x
- Gaut, B. S., Diez, C. M., and Morrell, P. L. (2015). Genomics and the contrasting dynamics of annual and perennial domestication. *Trends Genet.* 31, 709–719. doi: 10.1016/j.tig.2015.10.002
- Gaut, B. S., Seymour, D. K., Liu, Q., and Zhou, Y. (2018). Demography and its effects on genomic variation in crop domestication. *Nat. Plants* 4, 512–520. doi: 10.1038/s41477-018-0210-1
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: 10.1038/s41586-020-2649-2
- Herrera, J. C., Combes, M. C., Cortina, H., Alvarado, G., and Lashermes, P. (2002). Gene introgression into *Coffea arabica* by way of triploid hybrids (*C. arabica* × *C. canephora*). *Heredity* 89, 488–494. doi: 10.1038/sj.hdy.6800171
- Huang, L., Wang, X., Dong, Y., Long, Y., Hao, C., Yan, L., et al. (2020). Resequencing 93 accessions of coffee unveils independent and parallel selection during *Coffea* species divergence. *Plant Mol. Biol.* 103, 51–61. doi: 10.1007/s11103-020-00974-4
- Hufford, M. B., Lubinsky, P., Pyhäjärvi, T., Devengeno, M. T., Ellstrand, N. C., and Ross-Ibarra, J. (2013). The genomic signature of crop-wild introgression in maize. *PLoS Genet.* 9, e1003477. doi: 10.1371/journal.pgen.1003477
- Jolliffe, I. T., and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 374, 20150202. doi: 10.1098/rsta.2015.0202
- Lashermes, P., Andrzejewski, S., Bertrand, B., Combes, M.-C., Dussert, S., Graziosi, G., et al. (2000). Molecular analysis of introgressive breeding in coffee (*Coffea arabica* L.). *Theor. Appl. Genet.* 100, 139–146. doi: 10.1007/s001220050019
- Lashermes, P., Combes, M.-C., Ansaldi, C., Gichuru, E., and Noir, S. (2011). Analysis of alien introgression in coffee tree (*Coffea arabica* L.). *Mol. Breed.* 27, 223–232. doi: 10.1007/s11032-010-9424-2
- McKinney, W. (2010). “Data structures for statistical computing in Python”, in *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, eds. S. van der Walt and J. Millman (Austin, TX: SciPy), 56–61. doi: 10.25080/Majora-92bf1922-00a
- Merot-L’anthoene, V., Tournéize, R., Darracq, O., Rattina, V., Lepelley, M., Bellanger, L., et al. (2019). Development and evaluation of a genome-wide Coffee 8.5K SNP array and its application for high-density genetic mapping and for investigating the origin of *Coffea arabica* L. *Plant Biotechnol. J.* 17, 1418–1430. doi: 10.1111/pbi.13066
- National Center for Biotechnology Information (2024). *Coffea arabica* genome assembly *Coffea arabica* ET-39 v0.6, assembly accession GCA_036785775.1. NCBI Datasets. Available online at: https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_036785775.1/ (Accessed May 28, 2026).
- Ngure, G. M., and Watanabe, K. N. (2024). Coffee sustainability: leveraging collaborative breeding for variety improvement. *Front. Sustain. Food Syst.* 8, 1431849. doi: 10.3389/fsufs.2024.1431849
- Qian, L., Qian, W., and Snowdon, R. J. (2014). Sub-genomic selection patterns as a signature of breeding in the allopolyploid *Brassica napus* genome. *BMC Genomics* 15, 1170. doi: 10.1186/1471-2164-15-1170
- Rodrigues, A. S. B., Silva, D. N., Várzea, V., Paulo, O. S., and Batista, D. (2022). Worldwide population structure of the coffee rust fungus *Hemileia vastatrix* is strongly shaped by local adaptation and breeding history. *Phytopathology* 112, 1998–2011. doi: 10.1094/PHYTO-09-21-0376-R
- Salojärvi, J. (2023). The genome and population genomics of allopolyploid *Coffea arabica* reveal the diversification history of modern coffee cultivars. *Dryad*. doi: 10.5061/dryad.qnk98sfpt. Dataset.
- Salojärvi, J., Rambani, A., Yu, Z., Guyot, R., Strickler, S., Lepelley, M., et al. (2024). The genome and population genomics of allopolyploid *Coffea arabica* reveal the diversification history of modern coffee cultivars. *Nat. Genet.* 56, 721–731. doi: 10.1038/s41588-024-01695-w
- Santos, M. de L., Resende, M. L. V. de, Alves, G. S. C., Huguet-Tapia, J. C., Resende, M. F. R. de J., and Brawner, J. T. (2022). Genome-wide identification, characterization, and comparative analysis of NLR resistance genes in *Coffea* spp. *Front. Plant Sci.* 13, 868581. doi: 10.3389/fpls.2022.868581
- Scalabrín, S., Magris, G., Liva, M., Vitulo, N., Vidotto, M., Scaglione, D., et al. (2024). A chromosome-scale assembly reveals chromosomal aberrations and exchanges generating genetic diversity in *Coffea arabica* germplasm. *Nat. Commun.* 15, 463. doi: 10.1038/s41467-023-44449-8
- Scalabrín, S., Toniutti, L., Di Gasparo, G., Scaglione, D., Magris, G., Vidotto, M., et al. (2020). A single polyploidization event at the origin of the tetraploid genome of *Coffea arabica* is responsible for the extremely low genetic variation in wild and cultivated germplasm. *Sci. Rep.* 10, 4642. doi: 10.1038/s41598-020-61216-7
- Sera, T., Sera, G. H., Fazuoli, L. C., and Bettencourt, A. J. (2013). IPR 99 - Dwarf arabica coffee cultivar resistant to coffee ringspot virus. *Crop Breed. Appl. Biotechnol.* 13, 91–94. doi: 10.1590/S1984-70332013000100011
- Session, A. M. (2024). Allopolyploid subgenome identification and implications for evolutionary analysis. *Trends Genet.* 40, 621–631. doi: 10.1016/j.tig.2024.03.008
- Silva, D. N., Várzea, V., Paulo, O. S., and Batista, D. (2018). Population genomic footprints of host adaptation, introgression and recombination in coffee leaf rust. *Mol. Plant Pathol.* 19, 1742–1753. doi: 10.1111/mpm.12657

- Sotero, D., and Domingues, D. (2025). Dataset 1: Genomic assemblies and annotations of *Coffea* species and subgenomes. *Zenodo*. doi: 10.5281/zenodo.17250347. Dataset.
- Stephens, C., Hammond-Kosack, K. E., and Kanyuka, K. (2022). WAKsing plant immunity, waning diseases. *J. Exp. Bot.* 73, 22–37. doi: 10.1093/jxb/erab422
- Van der Vossen, H., Bertrand, B., and Charrier, A. (2015). Next generation variety development for sustainable production of arabica coffee (*Coffea arabica* L.): a review. *Euphytica* 204, 243–256. doi: 10.1007/s10681-015-1398-z
- Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Trans. Math. Softw.* 11, 37–57. doi: 10.1145/3147.3165
- Wang, Z., Yang, J., Cheng, F., Li, P., Xin, X., Wang, W., et al. (2022). Subgenome dominance and its evolutionary implications in crop domestication and breeding. *Hortic. Res.* 9, uhac090. doi: 10.1093/hr/uhac090
- Wu, Y., and Zhou, J. (2013). Receptor-like kinases in plant innate immunity. *J. Integr. Plant Biol.* 55, 1271–1286. doi: 10.1111/jipb.12123