

Graph-based pan-genome reveals structural and functional diversity across oil palm domestication gradients

Redi Aditama

IPB University

Heri Adriwan Siregar

IPB University

Zulfikar Achmad Tanjung

PT SMART Tbk

Diny Dinarti

IPB University

Sintho Wahyuning Ardie

IPB University

Willy Bayuardi Suwarno

IPB University

Tony Liwang

PT SMART Tbk

Edy Suprianto

Indonesian Oil Palm Research Institute (IOPRI)

Hernan Mauricio Romero

PT SMART Tbk

Condro Utomo

PT SMART Tbk

Sudarsono Sudarsono

`sudarsono_agh@apps.ipb.ac.id`

IPB University

Research Article

Keywords: Copy-number variation, Oil palm genomics, Presence–absence variation, Resistance genes, Structural variation

Posted Date: June 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8879446/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Plant Biology on June 25th, 2026.
See the published version at <https://doi.org/10.1186/s12870-026-09320-0>.

Graph-based pan-genome reveals structural and functional diversity across oil palm domestication gradients

Redi Aditama^{1,2}, Heri Adriwan Siregar^{1,3}, Zulfikar Achmad Tanjung², Diny Dinarti¹, Sintho Wahyuning Ardie¹, Willy Bayuardi Suwarno¹, Tony Liwang², Edy Suprianto³, Hernan Mauricio Romero², Condro Utomo², Sudarsono Sudarsono^{1,*}

¹Department of Agronomy and Horticulture, Faculty of Agriculture, IPB University, Bogor 16680, West Java, Indonesia

²Department of Biotechnology, Plant Production and Biotechnology Division, PT SMART Tbk, Bogor 16810, West Java, Indonesia

³Division of Oil Palm Breeding, Indonesian Oil Palm Research Institute (IOPRI), Medan 20158, North Sumatra, Indonesia

Running title:

Graph pan-genome reveals oil palm diversity...

*Corresponding author:

Sudarsono Sudarsono - Plant Molecular Biology Laboratory, Department of Agronomy and Horticulture, Faculty of Agriculture, IPB University, Jl. Meranti – IPB Dramaga Campus, Bogor 16680, West Java, Indonesia (Email: sudarsono_agh@apps.ipb.ac.id).

ORCID ID:

Redi Aditama – <https://orcid.org/0000-0003-2216-8798>

Heri Adriwan Siregar – <https://orcid.org/0009-0008-1116-3717>

Zulfikar Achmad Tanjung – <https://orcid.org/0009-0009-0820-6517>

Diny Dinarti – <https://orcid.org/0000-0002-7939-5476>

Sintho Wahyuning Ardie – <https://orcid.org/0000-0003-0563-1373>

Willy Bayuardi Suwarno – <https://orcid.org/0000-0002-7189-7906>

Tony Liwang – <https://orcid.org/0009-0005-1901-6410>

Edy Suprianto – <https://orcid.org/0000-0003-3395-8621>

Hernan Mauricio Romero – <https://orcid.org/0000-0002-0760-2296>

Condro Utomo – <https://orcid.org/0000-0002-8845-9195>

Sudarsono Sudarsono – <https://orcid.org/0000-0003-4688-5628>

Graph-based pan-genome reveals structural and functional diversity across oil palm domestication gradients

Abstract

Background Oil palm (*Elaeis guineensis* Jacq.), the world's most land-efficient oil crop, underpins global vegetable oil supply yet faces mounting constraints from limited expansion, climate stress, and disease pressure. These challenges highlight the urgent need for genomic resources that capture species-wide diversity to support sustainable improvement. While recent reference assemblies have advanced trait discovery, single linear genomes fail to represent the full spectrum of structural and gene-content variation, limiting resolution of agronomic alleles.

Results Here, we constructed a graph-based pan-genome from 30 diverse oil palm assemblies representing wild, semi-domesticated, and commercial accessions. We characterized structural variants, gene presence–absence variation, and copy-number gains, with focusing on functional stratification and resistance gene dynamics. The graph-based pan-genome revealed extensive structural and gene-content variation, including a large conserved core, complemented by shell and unique fractions enriched or biased toward regulatory, stress-responsive, and defense-related functions. Structural variation and duplication-derived copy-number gains contributed substantially to gene-content diversity, with semi-domesticated accessions exhibiting the greatest variability. Resistance gene repertoires showed contrasting patterns: receptor-like kinases remained comparatively stable, whereas the CNL subclass of NLR genes contributed disproportionately to shell-genome variation and duplication-associated turnover.

Conclusions This graph-based pan-genome provides a curated multi-assembly reference and comparative framework for oil palm genomics. By capturing structural variants, gene-content variations, copy-number gains, and resistance gene dynamics across domestication gradients, it establishes a foundation for future pan-GWAS analysis, functional genomics, and molecular breeding strategies aimed at improving resilience and productivity in this globally important crop.

Keywords:

Copy-number variation; Oil palm genomics; Presence–absence variation; Resistance genes; Structural variation

Introduction

Oil palm (*Elaeis guineensis* Jacq.) is widely recognized as the most land-efficient oil crop, producing several-fold more oil per hectare than soybean, rapeseed, or sunflower. Consequently, oil palm contributes disproportionately to global vegetable oil production [1–3]. As more oil palm plantation expansion is increasingly constrained by stringent no-deforestation regulations, yield gains must be achieved on existing acreages to meet future palm oil demand [4]. At the same time, intensifying biotic and abiotic pressures—most notably basal stem rot caused by *Ganoderma boninense* and rising heat and water stress under a warming climate—continue to erode productivity and plantation resilience of oil palm [5,6].

These challenges highlight the need for precision breeding methodology development and the strategic importance of high-quality, chromosome-scale reference genomes to support modern molecular breeding, including precise marker discovery, genome-wide association study (GWAS), and genomic selection. Recent oil palm genome assemblies have demonstrated the value of improved references by enhancing mapping resolution and revealing trait-relevant loci, exemplified by the identification of the *SHELL* MADS-box locus controlling fruit form and oil yield [7–9].

However, a single linear reference genome captures only a subset of species-wide variation and under-represents structural variants (SVs) and presence–absence variation (PAV), both of which can exert substantial phenotypic effects. Pan-genome studies across diverse crops routinely report sizeable fractions of genes affected by PAV and extensive copy-number changes that remain invisible in single-reference frameworks [10,11]. Such incompleteness propagates into downstream analyses through reference bias in read mapping and variant discovery, limiting the resolution of alleles associated with agronomic traits [12,13]. In contrast, pan-genome representations—particularly graph-based references constructed directly from multiple assemblies—explicitly encode alternative haplotypes and SVs as paths within a unified coordinate system, and this approach improves alignment accuracy and facilitates the analysis of complex variants [14,15]. Recent plant pan-genome studies demonstrate that key trait-associated sequences and SVs are accessible only when multiple assemblies are jointly represented [16–18]. These advances underscore a critical methodological gap between traditional single-reference frameworks and pan-genome-based approaches for capturing the full spectrum of genomic diversity.

Graph-based pan-genomes reduces reference bias and improves read alignment and genotyping of complex variants compared to linear references [19,20]. Recent methodological advances, including state-of-the-art builders such as Minigraph-Cactus, enable scalable graph construction directly from multiple assemblies while remaining compatible with downstream analysis workflows [14]. Independent evaluations in plants and other organisms further demonstrate that graph-aware mappers (e.g., vg/Giraffe) achieve higher accuracy and more reliable SV genotyping across diverse contexts, clarifying the trait-relevant variation that is difficult to resolve with a single reference [21,22].

In oil palm, domestication history has concentrated commercial planting material into a few pedigrees—notably the Deli Dura lineage traced back to four Bogor palms introduced in 1848—while large reservoirs of allelic diversity remain in semi-domesticated and wild populations across Africa [23–25]. Since the 1970s, frequent germplasm introductions and collection efforts were sought to broaden the selectable gene pools in response to narrowing

diversity of commercial materials [26]. Systematic characterization of wild African oil palm collections has revealed that substantial genetic variations are under-represented in Southeast Asian breeding pools [25,27]. In this context, constructing a pan-genome from a diverse panel—including wild, semi-domesticated, and commercial accessions— should provide a more reliable foundation for detecting lineage-restricted sequences and SVs of agronomic interest, as supported by emerging plant pan-genome studies [28].

Here, we assemble and analyze a graph-based pan-genome of oil palm from a selectively diverse panel of wild, semi-domesticated, and commercial accessions, anchored to the EG11 reference genome [8]. The study was designed to characterize structural variations, gene presence–absence variations, and duplication-derived copy-number gains across the wild, semi-domesticated, and commercial oil palm accessions; assess functional stratification among core, shell, and unique gene components; and examine the distribution, composition, and copy-number dynamics of resistance genes across accession groups. Together, this work delivers a curated multi-assembly reference resource and a comparative framework for defining structural and gene-content variation in oil palm, providing a foundation for future pan-GWAS, functional genomics, and molecular breeding applications.

Results

High-Quality Assemblies Enable Robust Pan-Genome Construction

Assembly quality was evaluated for each of the 30 oil palm accessions before graph construction. Per-accession sequencing and assembly statistics, including sequencing depth, ONT read N50, assembly size, contig number, contig N50, BUSCO completeness, and repeat content, are provided in Supplementary Table S2. Assembly contiguity was comparable across wild, semi-domesticated, and commercial oil palm accessions, with no significant differences in contig N50 (Kruskal-Wallis $p = 0.460$; Supplementary Table S3; Supplementary Fig. S1A). Median values were similar among commercial (0.253 Mb), semi-domesticated (0.201 Mb), and wild (0.224 Mb) groups (Supplementary Table S3).

Contig N50 correlated positively with ONT read N50 (Spearman $\rho = 0.44$, $p = 0.0149$) and assembly repeat content ($\rho = 0.56$, $p = 0.00217$), and covariate-adjusted analysis showed that these associations were not driven solely by sequencing depth (Supplementary Table S3; Supplementary Fig. S1C, D). Gene-space completeness was consistently high across assemblies. BUSCO scores showed a modest group-level effect (Kruskal-Wallis $p = 0.028$; Supplementary Table S3), driven by greater variability in semi-domesticated accessions, while wild and commercial groups remained similarly high (Supplementary Fig. S1B). Nearly all assemblies exceeded ~97% BUSCO completeness, indicating stable gene recovery across accessions and minimizing risk that downstream pan-genome patterns were driven by systematics differences in assembly quality. Overall, these quality metrics supported the use of the 30 assemblies for graph-based pan-genome construction.

Graph-Based Pan-Genome Captures Species-Wide Structural Diversity

Using polished assemblies that passed contiguity, BUSCO completeness, and repeat-content quality assessments, we constructed a graph-based pan-genome representing wild, semi-domesticated, and commercial oil palm accessions from Africa and Southeast Asia (Figure 1A). SNP-based phylogenetic analysis revealed genetic relationships broadly consistent with fruit type, germplasm provenance, and breeding history, together with substantial variation in heterozygosity among accessions (Figure 1B). This pattern supports the inclusion of genetically diverse genomes in the graph-based pan-genome.

To further evaluate the genetic relationships among accessions, we performed PCA using the same set of 558,789 informative genome-wide SNPs used for phylogenetic reconstruction. The PCA of 30 *E. guineensis* accessions showed that PC1 and PC2 explained 18.94% and 5.58% of the total variation, respectively. Accessions were not separated into fully discrete wild, semi-domesticated, and commercial clusters, but instead formed a continuum of genetic variation across the panel. Wild and semi-domesticated accessions occupied a broad and partially overlapping genetic space, whereas commercial accessions were also dispersed. This pattern is consistent with the composition of the commercial category, which included elite *Dura* and *Pisifera* parental lines as well as commercial *Tenera* hybrids in the oil palm hybrid breeding system (Supplementary Figure S2). Therefore, the wild, semi-domesticated, and commercial categories were interpreted as provenance- and breeding-status groups supported by genetic relationships, rather than as strictly discrete genetic clusters.

Graph growth analysis showed a continued increase in node number with sequential genome additions, indicating continued structural expansion of the oil palm pan-genome graph at the current sampling depth rather than definitive gene-level openness (Figure 1C). To separately evaluate gene-level pan-genome behavior, we used randomized-based rarefaction of gene presence-absence profiles (Figure 3A). Across repeated genome-order permutations, the total gene set continued to increase with additional genomes, but the rate of new gene discovery declined at higher sampling depth. Heaps' law modelling yielded an estimated gamma value of 0.441, consistent with continued but slowing gene-level expansion. Core and shell gene sets stabilized earlier, whereas unique genes continued to accumulate, indicating that residual pan-genome growth is driven primarily by rare or accession-restricted genes rather than broad expansion of all gene categories. Graph summary statistics are provided in Supplementary Table S4, whereas gene category counts and validation metrics are provided in Supplementary Table S7.

The coverage extended across all 16 chromosomes, with broadly consistent depth profiles across regions (Figure 1D), indicating that the graph provides a comprehensive representation of oil palm genome structure. Summary statistics are provided in Supplementary Table S4. The resulting graph revealed extensive structural diversity across chromosomes, prompting a detailed analysis of variant classes and distributions.

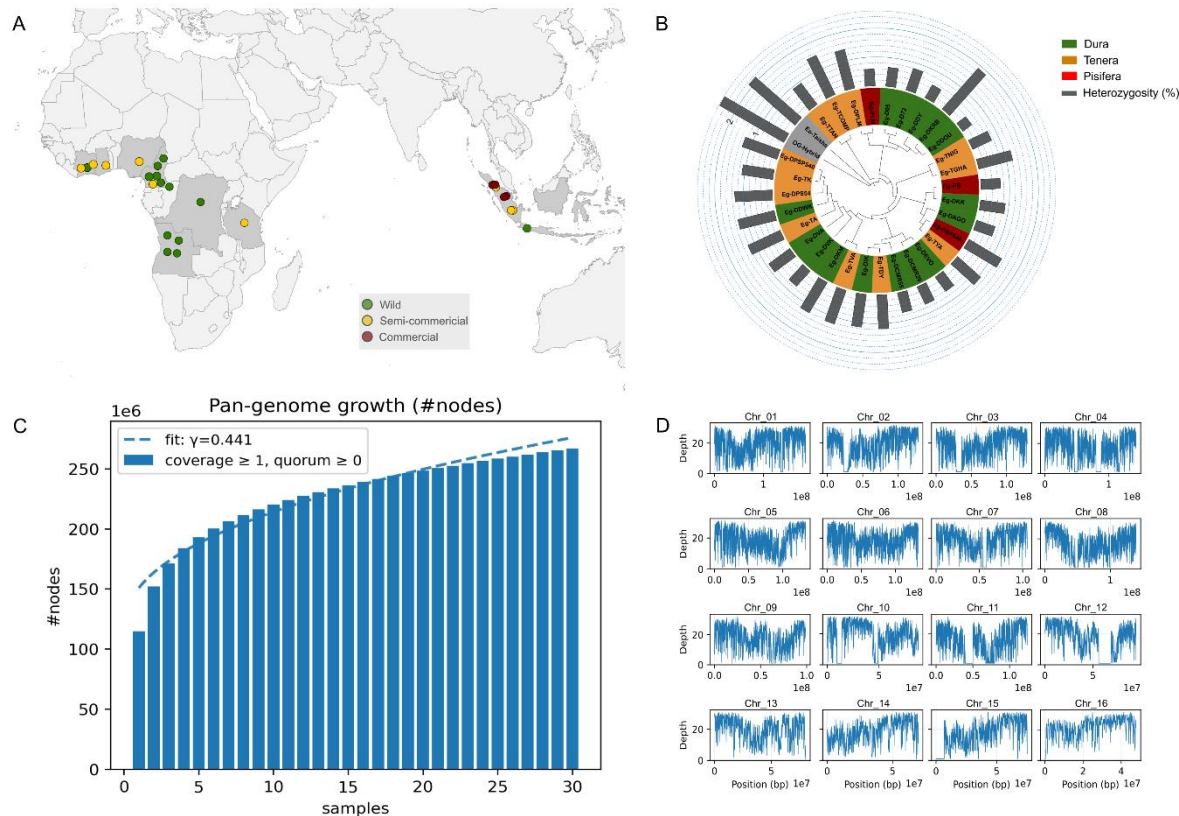


Figure 1. Construction and overview of the oil palm pan-genome graph. (A) Geographic origins of the oil palm accessions used for pan-genome construction, representing wild, semi-domesticated, and commercial groups across Africa and Southeast Asia. (B) SNP-based phylogenetic tree showing genetic relationships among accessions, with fruit type indicated by color and heterozygosity levels displayed as bar plots. (C) Pan-genome node growth curve showing the increase in the number of graph nodes with sequential addition of genomes, indicating continued structural expansion of the graph at the

current sampling depth. The dashed line indicates the fitted growth trend. (D) Graph coverage depth along the 16 oil palm chromosomes, demonstrating consistent representation of genomic regions across the pan-genome.

Extensive Structural Variation Across Oil Palm Chromosomes

Using the graph-based pan-genome, we characterized structural variation across 30 oil palm accessions, capturing simple insertions, deletions, and duplication-derived variants. Structural variants were operationally defined as SV50 events, corresponding to graph-derived variants with allele length ≥ 50 bp. Per-accession SV counts by variant classes are provided in Supplementary Table S5. Structural variants were distributed across all 16 chromosomes, and the graph contained long-range duplication-associated connections inferred from insertion-origin mappings (Figure 2A). Because these connections were derived from graph alignment and insertion-origin inference rather than orthogonal validation, they were interpreted conservatively as candidate long-range duplication-associated events rather than confirmed interchromosomal rearrangements.

Across the 16 chromosomes, we identified 318,744 SV50 events at the chromosome level (Supplementary Table S6). Insertions and deletions were the predominant classes, accounting for 135,362 (42.47%) and 137,205 (43.05%) events, respectively. Duplication-derived variants represented a smaller but substantial component of the SV landscape, including tandem duplications (2,967; 0.93%), proximal duplications (3,901; 1.22%), dispersed duplications (5,592; 1.75%), and putative interchromosomal duplications (33,717; 10.58%). These duplication-associated classes were defined operationally based on the physical distance between insertion sites and inferred origin loci, allowing local, regional, long-range intrachromosomal, and interchromosomal duplication-derived events to be distinguished. These results indicate that the oil palm pan-genome is dominated by insertion-deletion variation, while long-range and duplication-associated variants contribute additional structural complexity.

SV length distributions differed markedly among variant classes (Figure 2B). Insertions and deletions spanned broad size ranges, with deletions showing greater variability and more frequent large events. Duplication-related variants displayed more constrained distributions: tandem duplications were predominantly short, whereas proximal and dispersed duplications covered wider ranges with occasional large events. Interchromosomal duplications, though less frequent, also included sporadic long SVs.

Chromosome-level SV density varied substantially among chromosomes after normalization by chromosome length (Figure 2C; Supplementary Table S6). Because raw chromosome-level SV counts can be influenced by chromosome size, SV density was normalized by chromosome length. The highest total SV50 densities were observed on Chr_01, Chr_07, Chr_13, Chr_06, and Chr_16, whereas Chr_11, Chr_12, Chr_08, Chr_10, and Chr_04 showed lower densities. Collectively, these patterns reveal extensive structural variation spanning multiple classes and size ranges, providing a foundation for downstream analyses of gene-level presence-absence variation and duplication-derived copy-number gains.

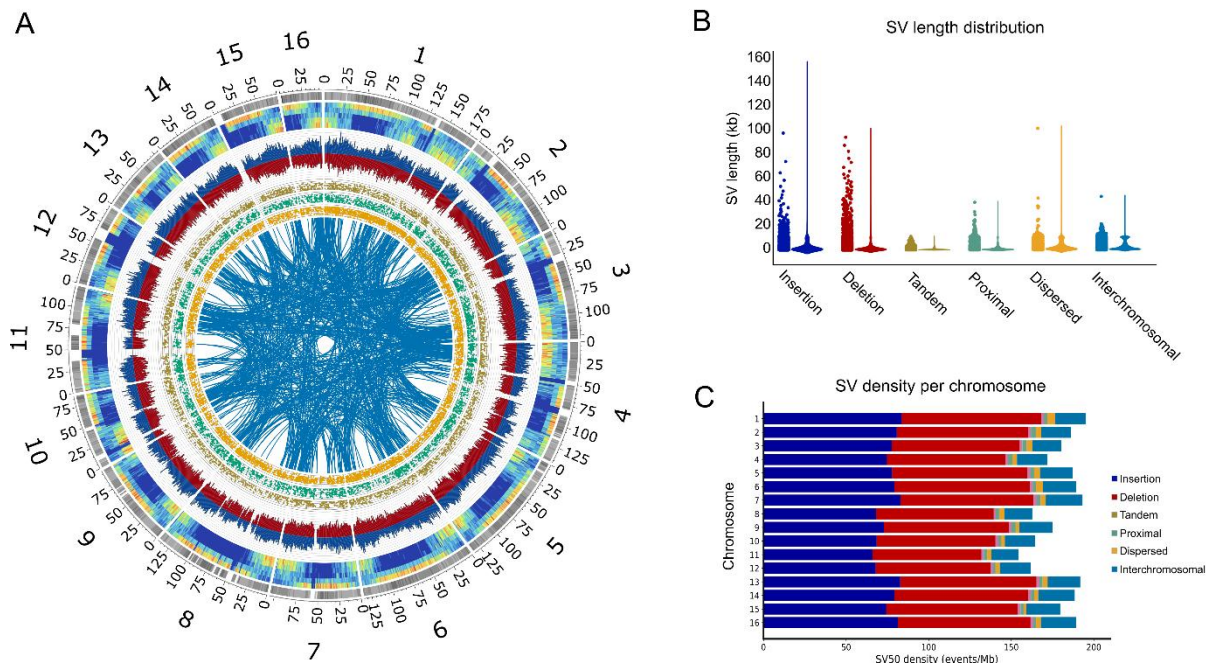


Figure 2. (A) Circos representation of structural variation and repeat-associated features across the 16 oil palm chromosomes, shown in 1-Mb windows. From outer to inner tracks: GC content; copy number of Copia, Gypsy, and LINE retrotransposons; abundance of low-complexity regions and simple sequence repeats; histograms of insertion and deletion events; histograms of tandem, proximal, and dispersed duplications; and links representing interchromosomal duplications inferred from graph-based insertion-origin mappings. (B) Length distributions of structural variant classes, including insertions, deletions, and duplication-derived variants, illustrating distinct size profiles among variant types. (C) Chromosome-level SV density, calculated as the number of SV50 events per megabase, showing normalized variation in SV burden and class composition across chromosomes. Per-accession SV counts are provided in Supplementary Table S5, and chromosome-level SV counts and densities normalized by chromosome length are provided in Supplementary Table S6.

Presence–Absence Variation Defines Core, Shell, and Unique Gene Sets

To assess gene content diversity across oil palm accessions, we analyzed presence–absence variation (PAV) within the graph-based pan-genome. Rarefaction analysis revealed a steady increase in total gene number with successive genome additions, approaching saturation at higher sampling depths (Figure 3A). Core and shell gene sets stabilized early, whereas unique genes continued to accumulate, indicating persistent diversity despite diminishing rates of novel discovery.

Gene occupancy distributions displayed a characteristic U-shaped profile (Figure 3B). Of 33,716 genes in the filtered PAV matrix, 25,531 genes (75.72%) were classified as core, 7,552 genes (22.40%) as shell, and 633 genes (1.88%) as unique (Supplementary Table S7). Thus, the oil palm pan-genome is dominated by a large, conserved gene complement, accompanied by a substantial shell component and a smaller accession-restricted unique fraction. This pattern indicates that gene-level pan-genome expansion is component-specific rather than uniform across categories. The conserved and moderately variable components were largely stabilized at the present sampling depth, whereas continued accumulation was concentrated in the unique fraction. Thus, gene-level expansion in this dataset is best interpreted as continued discovery of rare or accession-restricted genes rather than ongoing expansion of the core or shell components.

Assembly-based validation using gene-span alignments confirmed the robustness of SV-derived PAV inference. Core genes showed near-perfect concordance with assembly-based presence calls and negligible false absence rates. Shell genes exhibited intermediate concordance, consistent with greater uncertainty expected for variability present genes. Unique genes were conservatively inferred, with a low false presence rate of ~ 3.4% (Supplementary Table S7). Together, these validation results support the accuracy of category-specific occupancy profiles while indicating that interpretation of shell genes should remain comparatively cautious.

The PAV matrix also incorporated insertion-derived nonreference loci that passed stringent ORF integrity and transposable element filtering. Although most exhibited rare or private occupancy, a subset was detected across multiple accessions, indicating they are not assembly artifacts. Functional annotation revealed enrichment for defense-related and resistance-associated functions within this nonreference fraction, consistent with the structural and evolutionary dynamism of plant defense loci.

Hierarchical clustering of the binary PAV matrix revealed structured patterns of gene content variation across samples (Figure 3C). Accessions broadly clustered by domestication status, with wild, semi-domesticated, and commercial genomes showing distinct yet overlapping profiles. Groupwise comparisons highlighted substantial overlap among all three groups, while smaller subsets of genes were restricted to individual groups or specific group combinations (Figure 3D). Collectively, these results demonstrate that the oil palm pan-genome consists of a large, conserved core complemented by a variable gene pool differentially distributed among accessions. Beyond binary presence-absence patterns, gene content diversity also arises from quantitative differences in copy number, prompting further analysis of duplication-derived CNV gains.

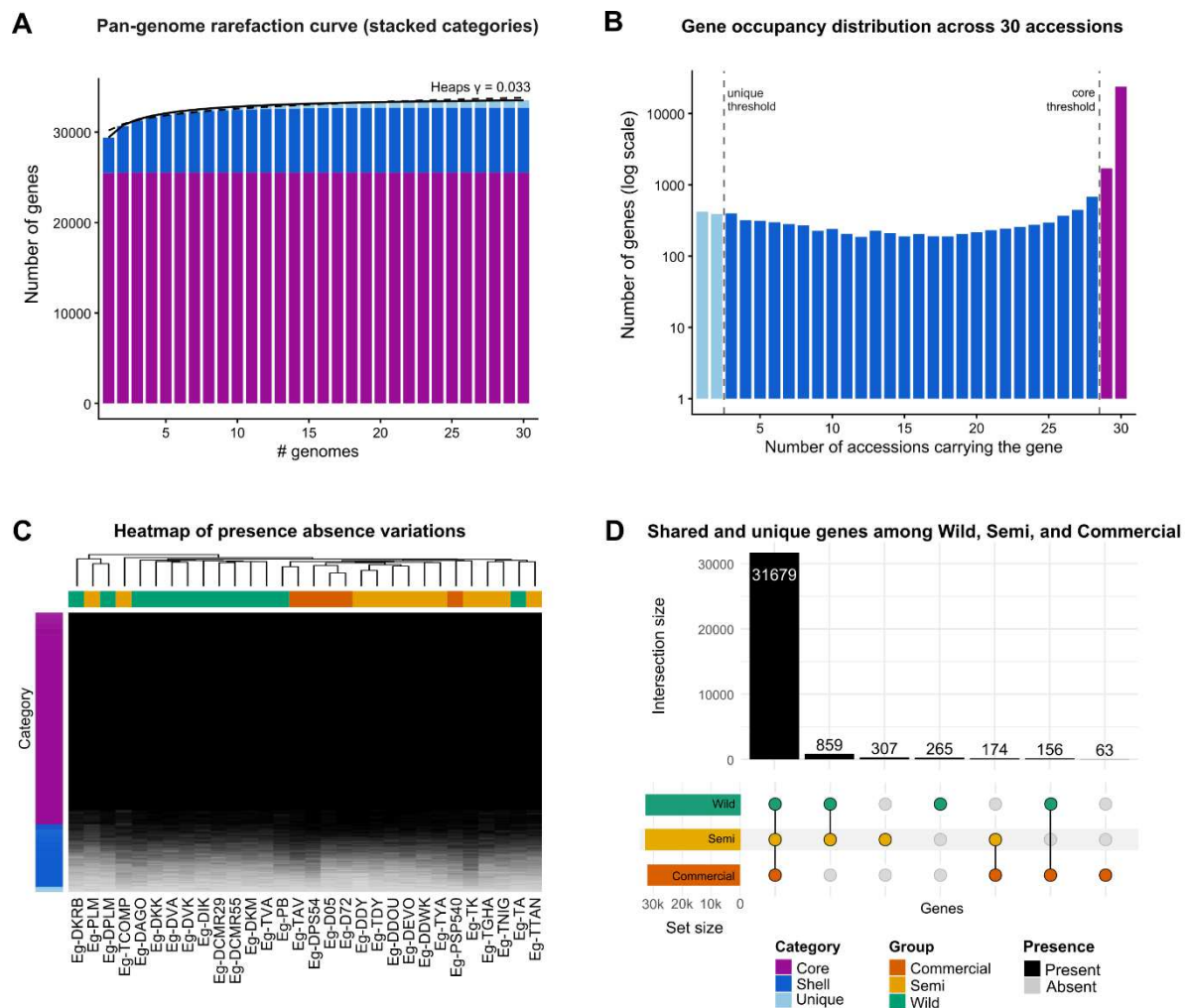


Figure 3. Presence–absence variation and pan-genome gene categories in oil palm. (A) Pan-genome rarefaction curve showing the accumulation of genes with increasing numbers of genomes, partitioned into core, shell, and unique gene categories. The fitted curve indicates a decreasing rate of novel gene discovery at higher sampling depth. (B) Gene occupancy distribution across the 30 accessions, highlighting a large core component, a variable shell component, and accession-restricted unique genes. (C) Hierarchical clustering of the binary gene presence–absence matrix across accessions and pan-genome categories. (D) UpSet plot showing shared and group-specific gene content among wild, semi-domesticated, and commercial accessions. Gene counts, category percentages, and assembly-based PAV validation metrics are provided in Supplementary Table S7.

Copy Number Variation Reveals Distinct Duplication Dynamics

To investigate structural variation at the gene level, we examined copy number variation (CNV) gains across 30 oil palm accessions. CNV here refers specifically to duplication-associated insertion events inferred from the pan-genome graph, rather than absolute copy number estimated from read depth. Overall, the number of genes showing CNV gain was broadly comparable among wild, semi-domesticated, and commercial groups (Kruskal-Wallis $H = 0.94$, $p = 0.63$; Figure 4A), indicating no significant differences in group medians. Semi-domesticated accessions, however, exhibited a wider spread of CNV values, reflecting increased variability rather than a systematic shift in overall burden. In contrast, wild and commercial groups showed more constrained distributions, suggesting relatively stable levels of gene copy-number gain.

Partitioning CNV gains by duplication class revealed distinct patterns across groups (Figure 4B). Tandem duplication contributed the largest fraction of CNV gains in all accessions, followed by interchromosomal duplications, while proximal and dispersed duplications accounted for smaller proportions. Chi-square tests confirmed significant differences in gain distributions among domestication groups for all duplication classes (tandem: $\chi^2 = 125.7$, $p < 1 \times 10^{-27}$; proximal: $\chi^2 = 33.2$, $p < 1 \times 10^{-7}$; dispersed: $\chi^2 = 28.1$, $p < 1 \times 10^{-6}$; interchromosomal: $\chi^2 = 115.6$, $p < 1 \times 10^{-25}$). Semi-domesticated accessions consistently showed higher proportions of proximal, dispersed, and interchromosomal duplications, whereas commercial accessions exhibited lower fractions across all classes (Supplementary Table S8).

CNV distributions across individual genomes revealed pronounced heterogeneity (Figure 4C). Many CNV gains were private or rare, present in only one or a few accessions, while others were common or fixed across multiple genomes. This spectrum of frequencies indicates that gene copy number expansion in oil palm includes both lineage-specific events and broadly shared duplications. Together, these results highlight the dynamic role of duplication-derived CNVs in shaping gene content diversity, with semi-domesticated accessions showing the greatest variability. To understand the biological implications of these variations, we next examined functional enrichment across core, shell, and unique gene sets.

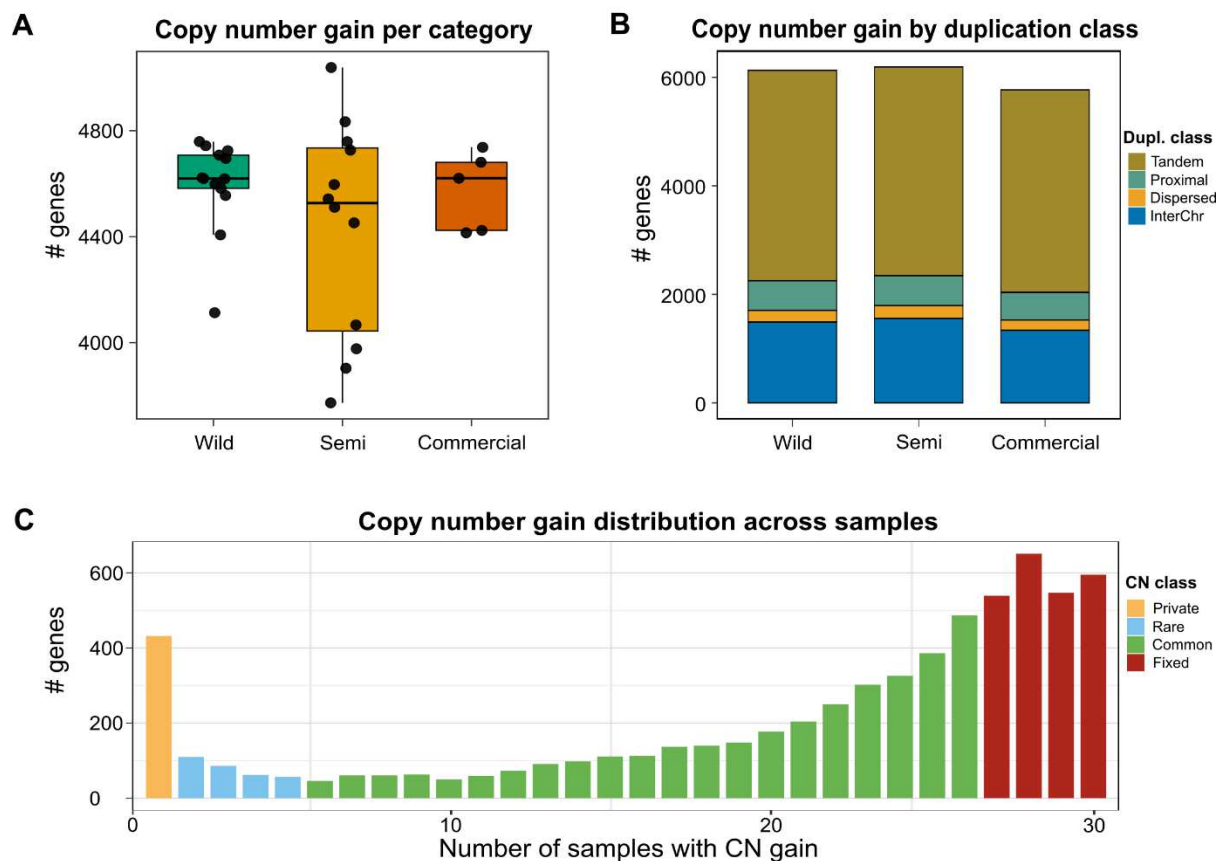


Figure 4. (A) Distribution of genes showing SV-derived copy number gain in wild, semi-domesticated, and commercial accessions. Boxplots summarize group-level variation, with individual accessions shown as points. (B) Composition of copy number gains by duplication class, including tandem, proximal, dispersed, and interchromosomal duplications, shown for each accession group. (C) Distribution of copy number gains across samples, illustrating the number of genes exhibiting private, rare, common, or fixed copy number gains as a function of the number of accessions carrying the gain.

Accession-level and group-level CNV summaries by duplication class are provided in Supplementary Table S8.

Functional Stratification of Core, Shell, and Unique Genes

Functional enrichment analyses revealed a consistent stratification of biological roles across pan-genome components: core genes were primarily associated with conserved cellular and housekeeping functions, shell genes represented an intermediate and more variable functional compartment, and unique genes were enriched or biased toward regulatory, stress-responsive, and specialized metabolic functions (Figure 5; Supplementary Table S9). Enrichment analyses were performed for GO terms, eggNOG-derived COG categories, and KEGG pathways using contingency-based tests with Benjamini–Hochberg correction. Full statistics, including observed and expected counts, enrichment ratios, raw p-values, and FDR values, are provided in Supplementary Table S9. In total, 82 GO terms, 11 COG categories, and 6 KEGG pathways showed significant enrichment or depletion at $FDR < 0.05$. For unique genes, significant enrichment after correction was detected for selected GO terms and the COG category related to signal transduction mechanisms, whereas KEGG pathway-level signals for unique genes were interpreted as directional trends rather than statistically significant enrichments after FDR correction.

The GO, eggNOG-derived COG category, and KEGG/KO annotations supported the same overall conclusion. GO annotations showed that core genes were associated with essential cellular processes, including nucleic acid metabolism, protein metabolism, cellular organization, and regulation of gene expression, while shell and unique genes were more frequently linked to regulation, localization, stimulus response, signaling, transport, and development. Broad eggNOG-derived COG category profiles further supported this pattern, with core genes enriched for conserved information-processing and cellular maintenance functions, whereas unique genes showed significant enrichment for signal transduction mechanisms. KEGG enrichment was strongest for core and shell gene sets, with core genes associated with housekeeping pathways such as ribosomal components, cofactor biosynthesis, and primary metabolism, while shell genes showed intermediate enrichment across several metabolic pathways. Biologically, this functional partitioning suggests that the conserved core genome maintains essential cellular activity across oil palm accessions, whereas variable shell and unique components may contribute to ecological adaptation, stress responsiveness, and specialized metabolism. These functions are agronomically relevant because variation in regulatory, signalling, transport, and metabolic genes may underlie accession-specific responses to pathogens, environmental stress, nutrient availability, and traits associated with productivity and resilience in oil palm breeding.

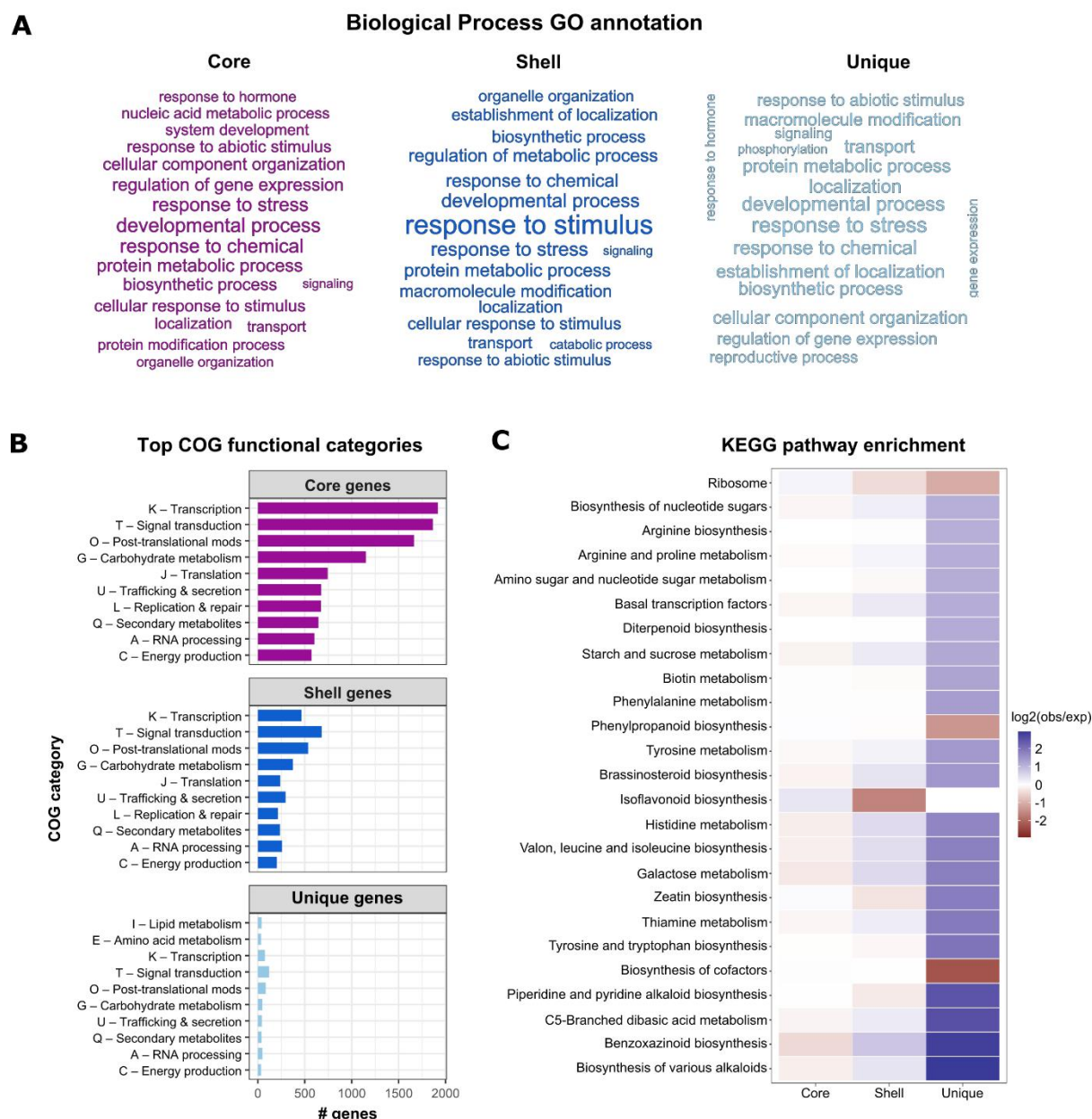


Figure 5. Functional stratification of core, shell, and unique genes. (A) Gene ontology (GO) biological process annotations for core, shell, and unique genes, showing distinct functional themes among pan-genome components. (B) Distribution of top eggNOG-derived COG functional categories across pan-genome classes, highlighting differences in functional composition between conserved and variable genes. (C) KEGG pathway enrichment across core, shell, and unique gene sets, shown as $\log_2(\text{observed}/\text{expected})$ values, illustrating contrasting enrichment and depletion patterns among pan-genome components. Full enrichment statistics, including observed and expected counts, enrichment ratios, raw p-values, and FDR values, are provided in Supplementary Table S9.

Resistance Gene Repertoires Show Contrasting Stability and Variability

We quantified resistance (R) gene repertoires across accessions by integrating presence-absence variation (PAV) with copy-number gains from duplication-associated insertion events. The combined measure of total R-gene repertoire size showed broadly comparable values among wild, semi-domesticated, and commercial accessions, with extensive overlap across groups and no significant difference in total counts (Kruskal-Wallis $p=0.29$; Figure 6A;

Supplementary Table S10). Thus, domestication category did not strongly differentiate overall R-gene repertoire size.

To assess how similar total repertoire sizes arise, we examined R-gene CNV burden separately. R-gene CNV burden, defined as the number of R-genes with copy-number gains per accession, overlapped across groups but was more variable in semi-domesticated accessions, spanning both the lowest and highest values (Figure 6B; Supplementary Table S10). Wild and commercial groups showed more constrained distributions. These results suggest greater heterogeneity in semi-domesticated materials, with CNV gains contributing variably rather than uniformly to R-gene repertoires.

Partitioning R-genes by pan-genome category revealed class-specific enrichment patterns (Figure 6C; Supplementary Table S11). Among nucleotide-binding leucine-rich repeat (NLR)-related classes, CNL genes, which represent coiled-coil NLRs and therefore a subclass of NLR genes, were significantly depleted from the core genome and enriched in the shell, reflecting pronounced presence-absence dynamics. By contrast, NL and TNL genes did not deviate significantly from genome-wide expectations. RLK genes showed the opposite trend to CNLs, being enriched in the core and depleted in the shell, consistent with stronger conservation of receptor-like kinase-mediated immune functions. Other classes, including RLP and MLO, also showed distributions broadly consistent with genome-wide expectations.

A stability-variability map summarizing class distributions reinforced these patterns (Figure 6D; Supplementary Table S12). RLK genes clustered toward the conserved end of the spectrum, driven by high core representation, whereas CNL genes shifted toward variability due to shell enrichment. NL genes also showed a high core fraction but did not differ significantly from genome-wide expectations, indicating that their distribution reflected the overall R-gene background rather than a class-specific bias. TNL, RLP, and MLO genes occupied intermediate positions without significant enrichment or depletion.

Enrichment analysis based on standardized residuals confirmed non-random associations between R-gene class and pan-genome category (Figure 6E; Supplementary Table S11). CNL genes were significantly depleted from the core and enriched in the shell, whereas RLK genes showed the opposite pattern. In contrast, NL and TNL genes followed genome-wide expectations, and RLP and MLO genes showed no significant deviation. These results indicate that the variable NLR-related signal was driven mainly by the CNL subclass rather than by all NLR-related genes uniformly.

Collectively, these findings demonstrate that R-gene classes differ markedly in their contribution to pan-genome variability. RLKs remain comparatively stable across oil palm lineages, whereas the CNL subclass of NLR genes contributes disproportionately to shell-genome variation. This contrasting stability and variability highlight the evolutionary dynamics of defense loci and their potential relevance for breeding strategies aimed at enhancing resilience against biotic stress.

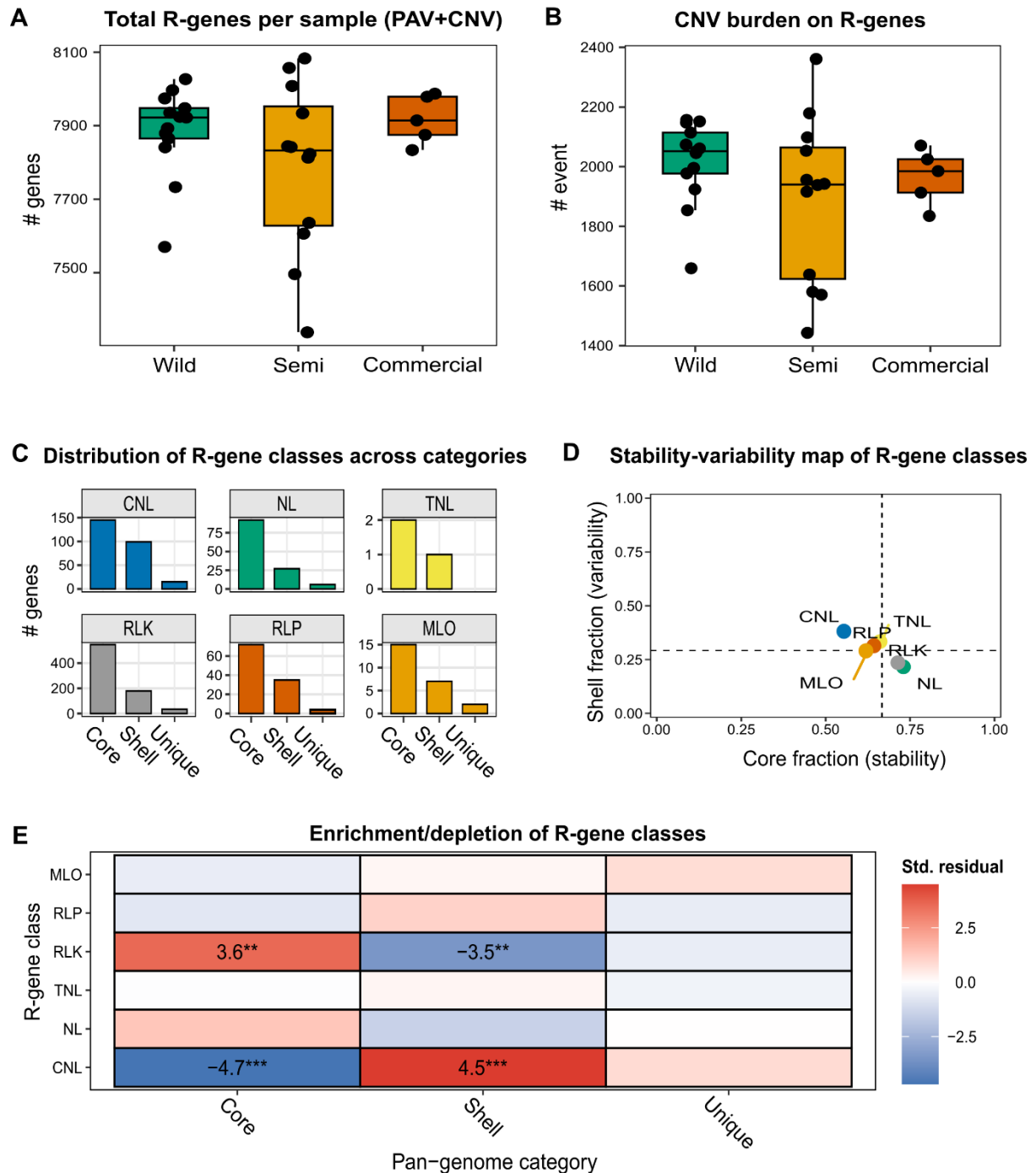


Figure 6. Distribution and pan-genome dynamics of resistance genes (R-genes). (A) Total number of R-genes per accession based on integrated presence-absence variation and copy number variation, for wild, semi-domesticated, and commercial groups (Supplementary Table S10). (B) SV-derived CNV burden affecting R-genes across accession groups (Supplementary Table S10). (C) Distribution of major R-gene classes across pan-genome categories (core, shell, and unique), with enrichment assessed using standardized residuals (Supplementary Table S11). CNL genes represent coiled-coil NLRs and therefore constitute a subclass of NLR genes. (D) Stability-variability map summarizing the proportion of core and shell genes for each R-gene class (Supplementary Table S12). Rightward displacement indicates increasing representation in the shell component (higher variability), whereas upward displacement indicates dominance in the core genome (greater stability). (E) Enrichment and depletion of R-gene classes across pan-genome categories based on standardized residuals, with positive and negative values indicating over- and under-representation, respectively (Supplementary Table S11).

Discussion

A central motivation for constructing a graph-based pan-genome is to overcome the limitations of a single linear reference in representing species-wide genomic diversity [14]. In the present study, the use of multiple de novo assemblies allowed structural and gene-content variation to be captured across wild, semi-domesticated, and commercial oil palm accessions. Because assembly contiguity and gene completeness were broadly comparable among groups, the observed variation is unlikely to reflect systematic differences in assembly quality alone. Instead, it highlights the extent of genomic diversity that becomes visible when oil palm genomes are represented jointly rather than projected onto a single reference.

The growth patterns observed in this study further indicate that pan-genome expansion is scale-dependent. At the graph level, node number continued to increase with the addition of genomes, reflecting ongoing sequence-level structural expansion. Such expansion can include additional haplotypes, indels, rearrangement-associated segments, non-coding regions, repeats, and transposable element-associated sequences. Gene-level rarefaction, however, showed a more constrained pattern: the core genome stabilized early, while the variable fraction continued to accumulate more slowly, mainly through rare or accession-restricted genes. This distinction is important because graph nodes and annotated genes represent different biological units and should not be interpreted as equivalent measures of pan-genome openness.

This scale-dependent pattern is consistent with plant pan-genome studies showing that variable genome compartments are shaped not only by protein-coding gene gain and loss, but also by structural variation, transposable element activity, deletions, rearrangements, and repetitive sequences[18,29,30]. Therefore, the higher growth parameter observed at the graph level likely reflects continued discovery of repeat-rich and TE-associated sequence variation rather than proportional discovery of new protein-coding genes. The oil palm pan-genome should therefore not be described as uniformly open across all components; instead, it shows persistent structural graph expansion alongside more constrained gene-level expansion driven mainly by rare or accession-restricted genes.

Within this framework, gene content is defined by structurally supported presence–absence patterns rather than reference annotation alone. Beyond the conserved reference gene set, the pan-genome captures gene-sized loci introduced through insertion events that are absent from the reference but recur across multiple assemblies. These loci contribute to the shell and unique components, reflecting lineage-restricted gene content systematically missed by single-reference approaches. By explicitly representing alternative haplotypes and structural variants within a unified coordinate system, the graph-based pan-genome enables genome-wide representation of insertions, deletions, and duplication-derived variants across all chromosomes [19]. In the oil palm pan-genome, heterogeneous structural variant distributions and interchromosomal duplication-associated connections inferred from graph-based insertion-origin mappings highlight graph-derived signals of long-range structural complexity that are difficult to represent with a single linear reference. Comparable patterns have been reported in other plant pan-genomes, including maize, wheat, and Brassica, where substantial fractions of functionally relevant variation are associated with structural and presence–absence variation under-represented in linear references [30–32]. In this context, the broadly uniform graph coverage observed across chromosomes indicates that the constructed

pan-genome provides a balanced representation of oil palm genomic regions, reducing reference bias and supporting downstream comparative analyses.

The domestication history of oil palm [33] provides important context for interpreting the observed pan-genome structure. Commercial planting material derives largely from a narrow genetic base, most notably the Deli Dura lineage, whereas wild and semi-domesticated accessions retain substantially greater diversity originating from West and Central Africa [9,34,35]. Consistent with this history, patterns of presence–absence variation (PAV) and copy number variation (CNV) reveal differences in genome architecture among accession groups. While a large core genome is shared across all accessions, shell and unique fractions exhibit structured variation that broadly corresponds to domestication status. Semi-domesticated accessions show greater heterogeneity in both PAV and CNV burden, reflecting admixture between wild and cultivated backgrounds and incomplete fixation during early domestication and breeding [34,36]. Similarly, the dispersion of commercial accessions in the SNP-based PCA is biologically expected because this group includes elite *Dura* female parents, elite *Pisifera* male parents, and commercial *Tenera* hybrids, which represent different components of the oil palm hybrid breeding system rather than a single homogenous genetic population.

The recurrent observation that semi-domesticated accessions showed the greatest variability across PAV, CNV, and R-gene analyses likely reflects their intermediate and heterogeneous breeding status. Unlike wild accessions, which preserve geographically structured ancestral variation, and commercial accessions, which have undergone stronger selection and narrowing through elite breeding, semi-domesticated materials often represent wild-derived germplasm that has experienced limited crossing, maintenance, or preliminary selection without full commercial fixation. This history can generate mosaic genomes that retain substantial ancestral diversity while also carrying structural variants introduced, reshuffled, or retained during early breeding and germplasm evaluation. Consequently, semi-domesticated accessions may show broader dispersion in gene-content profiles, copy-number gain burden, and defense-gene composition, not because they represent a single coherent domestication stage, but because they capture a transitional pool between wild diversity and commercial selection.

Structural variation further supports this pattern. Insertions and deletions dominate the SV landscape across all groups, but duplication-derived variants also contribute to genomic diversification. The coexistence of private, rare, and globally shared CNVs suggests that gene copy number growth in oil palm is shaped by both lineage-specific processes and conserved structural features. Similar trends have been reported in other domesticated plants, where CNVs reflect domestication gradients and contribute to phenotypic diversification [37–39]. In this context, although group-level differences in copy-number gain were modest, the broader dispersion observed in semi-domesticated accessions points to heterogeneous genomic outcomes during intermediate stages of domestication. Rather than reflecting a uniform directional shift, these patterns are consistent with incomplete fixation and admixture between wild and cultivated backgrounds, resulting in variable retention and expansion of gene copies across lineages and indicating that domestication altered oil palm genome architecture in diverse ways rather than uniformly reducing genomic diversity.

Functional stratification further supports the biological relevance of the oil palm pan-genome. Core genes were primarily associated with conserved cellular and housekeeping functions, whereas shell and unique genes showed stronger representation of regulatory, stress-

responsive, signalling, transport-related, and specialized metabolic functions. This pattern was supported by complementary GO, eggNOG-derived COG category, and KEGG/KO enrichment analyses. We retained eggNOG-derived COG categories as broad functional descriptors because they provide a compact high-level summary of functional composition across pan-genome components. However, because COG categories are coarse and legacy-derived, they were interpreted only as supporting evidence, with finer biological interpretation based primarily on GO and KEGG/KO annotations. Together, these results suggest that non-core genome components may contribute to ecological adaptation and accession-specific functional diversity, including responses to environmental stress, pathogen pressure, nutrient availability, and traits relevant to oil palm productivity and resilience.

Similar functional polarization between conserved and variable fractions has been reported across multiple plant pan-genomes, where non-core genes contribute disproportionately to adaptive, regulatory, and lineage-specific functions [31,40–42]. Collectively, these findings indicate that functional diversification in the oil palm pan-genome is driven primarily by structurally variable gene content, whereas conserved components maintain essential cellular functions.

Resistance genes illustrate how gene-content variation driven by SV-associated PAVs and CNVs contributes to functional diversity. Although the overall size of the R-gene repertoire is broadly comparable among wild, semi-domesticated, and commercial accessions, the structural mechanisms underlying this diversity vary across groups. Semi-domesticated accessions show greater dispersion in both PAV and CNV affecting R-gene loci, reflecting heterogeneous combinations of ancestral variation and incomplete fixation during early domestication and breeding. This pattern suggests that resistance gene architectures remain dynamic during intermediate stages of domestication rather than undergoing uniform erosion or expansion.

Beyond group-level comparisons, R-gene classes occupy distinct positions along a stability–variability spectrum in the pan-genome. Receptor-like kinases (RLKs) were predominantly assigned to the core genome and significantly enriched in the core component, indicating strong conservation across accessions and functional constraints on receptor-mediated immune signaling. RLP genes also showed substantial core representation, although their distribution did not deviate significantly from genome-wide expectations. Such conservation aligns with the central role of surface-associated receptors in basal immune perception and signal transduction, which are often under strong purifying selection due to their involvement in essential defense pathways⁴⁴. In contrast, nucleotide-binding leucine-rich repeat (NLR) genes, particularly the CNL subclass, are overrepresented in the shell component, consistent with elevated structural turnover and copy-number modulation. These contrasting patterns mirror observations in other plant species, where intracellular NLR-type genes undergo rapid birth–death dynamics driven by host–pathogen coevolution [43,44].

Enrichment analyses further support class-specific trends by quantifying associations between R-gene classes and pan-genome components. RLKs were significantly enriched in the core genome and depleted from the shell, whereas CNL genes showed the opposite pattern, enriched in the shell and depleted from the core. Because CNLs are coiled-coil NLRs and therefore represent a subclass of NLR genes, this result indicates that the variable NLR-related signal is driven primarily by CNLs rather than by all NLR-related classes uniformly. In contrast, NL and TNL genes followed genome-wide expectations, and RLP and MLO genes did not show

significant enrichment or depletion. This pattern suggests that oil palm defense-gene variability is class-specific, with conserved receptor-like immune components coexisting with more dynamic intracellular immune receptor subclasses, consistent with comparative and population-scale analyses of plant immune repertoires [43–45].

These patterns connect pan-genome structure with potential breeding relevance. Variable genome compartments, including shell and unique genes as well as duplication-derived copy-number gains, were enriched or biased toward regulatory, stress-responsive, and resistance-related functions. This suggests that non-core components may contribute to ecological adaptation and accession-specific functional diversity, particularly in semi-domesticated materials that retain heterogeneous combinations of ancestral and breeding-associated variation. Although the phenotypic effects of these variants remain to be validated, they provide a valuable reservoir of candidate lineage-restricted alleles for future pan-GWAS, genomic selection, and breeding strategies aimed at enhancing resilience and productivity.

Some limitations of this study should be acknowledged. Although the accession panel includes wild, semi-domesticated, and commercial materials, it does not capture the full geographic and genetic diversity of *E. guineensis*. Populations from eastern Africa, including Tanzania, remain underrepresented despite their potential importance for understanding historical dispersal and regional differentiation [9,34,35]. Inclusion of additional African populations would likely expand the variable component of the genome.

The unequal sample size among germplasm groups should be considered when interpreting group-level comparisons. In particular, the commercial group included fewer accessions than the wild and semi-domesticated groups, reflecting the narrower representation of elite released materials in the present panel. Consequently, statistical comparisons involving commercial accessions have reduced power to detect subtle differences and may underestimate within-group diversity. We therefore interpreted group-level patterns conservatively, emphasizing broad trends in structural variation, gene-content variation, CNV burden, and R-gene dynamics rather than definitive domestication effects.

Another limitation is that the present study did not directly test associations between PAV, CNV, and phenotypic traits. Therefore, functional enrichment results and the observed enrichment of regulatory, stress-responsive, metabolic, and resistance-related genes in variable genome compartments should be interpreted as candidate functional signals rather than direct evidence of breeding-relevant effects. Although these patterns suggest that shell and unique genes may contribute to adaptive or agronomic variation, their utility for breeding will require validation through phenotype-linked analyses such as pan-GWAS, expression profiling, controlled stress or disease assays, and marker–trait association studies in larger breeding populations.

Finally, SV detection and copy-number gain inference remain influenced by assembly quality, repeat resolution, haplotype representation, and graph construction parameters. Although all assemblies showed high BUSCO completeness and broadly comparable contiguity, haplotype purging was not performed before graph construction. This approach preserved accession-specific structural and haplotype-associated variation, but residual alternative haplotigs in highly heterozygous or repetitive regions may inflate local sequence redundancy, graph node counts, rare or unique gene-content calls, or duplication-associated signals. Because tandem, proximal, dispersed, and interchromosomal classes were inferred from graph-derived insertion-

origin mappings, they should be interpreted as operational spatial categories rather than confirmed molecular mechanisms. Interchromosomal connections require particular caution because they may reflect genuine long-range duplication-derived variation but may also arise from assembly fragmentation, misjoins, repetitive-sequence alignment ambiguity, retained haplotigs, or graph construction artifacts. To reduce these effects, we used quality-controlled assemblies, graph-based variant detection, stringent filtering of insertion-derived gene candidates, chromosome-level SV normalization, and assembly-based validation of PAV calls. Nevertheless, future studies using more contiguous long-read assemblies, haplotype-resolved assemblies, optical mapping, Hi-C-assisted scaffolding, and phenotype-linked validation will be important to refine SV resolution, validate long-range duplication-associated signals, and assess the functional impact of pan-genome variation.

Despite these limitations, the graph-based pan-genome presented here provides a valuable reference framework for future oil palm genomics. By capturing structural, functional, and resistance gene variation across domestication gradients, it expands genomic analyses beyond nucleotide substitutions to encompass gene presence–absence variation, copy number gains, and haplotype-level diversity. Notably, SV-mediated gene presence–absence variation and duplication-driven copy number gains represent major sources of functional variation that remain systematically underrepresented in conventional SNP-based frameworks relying on linear references, including at resistance gene loci [20,32,46]. Although the phenotypic effects of these variants remain to be validated, the catalogue generated here provides a foundation for future pan-GWAS, marker development, genomic selection, and functional studies aimed at linking structural and gene-content variation with disease resistance, stress tolerance, yield-related traits, and breeding resilience in oil palm.

Methods

Plant materials and accession classification

The oil palm (*E. guineensis* Jacq.) panel analyzed in this study comprised 30 accessions representing three germplasm categories: wild, semi-domesticated, and commercial materials. Accessions were assigned to domestication categories based on provenance records, breeding history, germplasm status, and known use in oil palm improvement programs, rather than solely on clustering in the phylogenetic tree. The wild category included accessions derived from formal African germplasm exploration programs and one historically important Bogor Botanical Gardens-derived accession. Specifically, the wild materials included oil palms collected during the 2008 Cameroon exploration, conducted through collaboration between the Indonesian Palm Oil Board (IPOB) and the Institute of Agricultural Research for Development of Cameroon (IRAD), and the 2010 Angola exploration, conducted through collaboration between IPOB and the Malaysian Palm Oil Board together with the Instituto Nacional do Café (INCA), Angola. This category also included Eg-DBBG, a first-generation descendant of one of the four original oil palms introduced to and planted in the Bogor Botanical Gardens in 1848. These materials were considered wild or historically derived germplasm because they represent non-commercial or minimally improved genetic resources maintained primarily for conservation, genetic diversity, and breeding exploration.

Semi-domesticated accessions were defined as wild-derived germplasm materials that had undergone at least one controlled cross or breeding step and were maintained and evaluated for specific breeding-related traits. These materials had been incorporated into breeding prospecting or pre-breeding activities but had not yet completed the selection, testing, and release process required for commercial deployment. Thus, the semi-domesticated group represents an intermediate category between wild genetic resources and fully released commercial breeding materials.

Commercial accessions were defined as elite oil palm materials that had been commercially released or were directly used in commercial seed production and planting programs. This group included superior female parents (*Dura*), superior male parents (*Pisifera*), and commercial hybrid planting materials. These accessions represent advanced breeding products subjected to stronger human selection and commercial performance evaluation than the wild and semi-domesticated materials.

No new field collections were undertaken for this study. All plant materials were obtained from established germplasm collections, breeding populations, or commercial breeding resources maintained under institutional oversight by PT SMART Tbk and/or the Indonesian Oil Palm Research Institute (IOPRI). Plant materials were handled in accordance with applicable institutional and national regulations. Formal taxonomic identification as *Elaeis guineensis* Jacq. was conducted by Heri Adriwan Siregar at IOPRI based on germplasm documentation and morphological characterization and was further supported by whole-genome sequencing analyses performed in this study. The species investigated is not listed under the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). Detailed accession metadata, domestication category assignments, fruit types, provenance, germplasm source, public database accessions, and release status are provided in Supplementary Table S1.

Sample and sequencing

We sampled 30 oil palm accessions representing wild, semi-domesticated, and commercial germplasm categories and extracted genomic DNA from young leaf tissue using a modified cetyltrimethylammonium bromide (CTAB) protocol. Accession metadata, domestication classification, fruit type, provenance, sequencing information, and public database accession status are summarized in Supplementary Table S1. DNA quality was assessed with a NanoDrop 2000 (Thermo Scientific), integrity verified by agarose gel electrophoresis and an Agilent 4150 TapeStation, and concentrations measured with Qubit dsDNA HS Assay Kits.

For short-read sequencing, high-molecular-weight DNA was enzymatically fragmented to the target insert size, end-repaired, ligated to MGI-compatible barcode adapters, PCR-amplified, quality checked (TapeStation/Qubit), circularized into DNA nanoballs (DNBs), and sequenced on a DNBSEQ-G400 platform using 612-cycle paired-end chemistry (PE300). In parallel, long-read libraries were prepared from the same DNA by end-repair and dA-tailing to generate 5'-phosphorylated/3'-dA-tailed fragments, ligated to Oxford Nanopore adapters, quantified with Qubit, and sequenced on a GridION (Oxford Nanopore Technologies). Sequencing was monitored in MinKNOW, raw signals were base-called with Guppy in High Accuracy mode, and runs continued until per-sample yield targets were met.

Population analysis

Raw short-read sequences from 30 *E. guineensis* oil palm accessions, together with one *Elaeis oleifera* accession and one interspecific *E. oleifera* × *E. guineensis* hybrid accession, were quality-filtered and adapter-trimmed using fastp v0.20.1 [47]. Adapter sequences were detected automatically, and reads were filtered using default fastp thresholds, including a qualified base Phred score threshold of Q15, a maximum allowed proportion of low-quality bases of 40%, a maximum ambiguous-base count of 5, and a minimum read length of 15 bp after trimming. These settings were used to remove sequencing adapters and low-confidence bases while retaining sufficient genome-wide coverage for SNP calling. Cleaned reads from each accession were mapped to the EG11 oil palm reference genome [8] using BWA-MEM v0.7.18 [48] with default parameters. Alignments were processed into sorted BAM files, and bcftools mpileup was used for variant calling. The resulting VCF files from individual accessions were merged into a combined dataset using bcftools merge [49]

SNP filtering was performed using PLINK v2.00a3 [50] with the following criteria: minimum allele frequency (--maf 0.05), maximum proportion of missing data (--max-missing 0.1), and exclusion of indels. These thresholds were selected to retain common, well-supported SNPs suitable for phylogenetic and PCA analyses while reducing the influence of rare variants, excessive missingness, and alignment uncertainty around indels. After filtering, a total of 11,381,742 high-quality SNPs were retained. The filtered VCF was converted into FASTA format using VCFkit v0.2.9 [51], and a phylogenetic tree was constructed using the neighbor-joining (NJ) method implemented in MEGA X [52]. The tree was rooted using *Elaeis oleifera* as the outgroup, while the interspecific hybrid (*E. guineensis* × *E. oleifera*) was included as a comparative reference. The resulting tree was visualized and annotated with the Interactive Tree of Life (iTOL v6) web server [53].

To complement the SNP-based phylogenetic analysis, principal component analysis (PCA) was performed using the cleaned SNP alignment. For the PCA focused on the *E. guineensis*

domestication and breeding gradient, only 30 *E. guineensis* accessions were retained, whereas *E. oleifera* outgroup and the interspecific hybrid were excluded. After removing non-informative sites, 558,789 variable SNP sites were retained for PCA. Each SNP position was numerically encoded by assigning the major allele as 0 and any alternative allele as 1. Ambiguous bases, gaps, and missing calls were treated as missing data and imputed using the mean value for each SNP site. The resulting SNP matrix was standardized, and PCA was performed in Python using the scikit-learn package. PCA results were visualized by coloring accessions according to germplasm category and shaping points according to fruit type or breeding role (*Dura*, *Pisifera*, or *Tenera*). The PCA was used to visualize genetic relationships and the overall genetic continuum represented by the accession panel rather than to define discrete population-genetic clusters.

Genome assembly and polishing

Long-read data were quality-controlled with fastplong v0.1.1 [47] to remove low-quality reads and sequencing artifacts before assembly. Reads passing the quality-filtering step were assembled using Flye v2.9.3 [54] with an expected genome size of 1.8 Gb and default parameters. The use of quality-filtered long reads was intended to reduce the contribution of low-confidence ONT reads to assembly graph construction while preserving long-range information required for contig assembly. Base-level accuracy was improved by polishing each assembly with quality-filtered short reads processed with fastp v0.20.1 [47]. Reads were aligned to the assemblies using BWA-MEM v0.7.18 [48], and iterative corrections were performed with Pilon v1.24 [55]. Short-read polishing was used to reduce residual ONT-associated base errors and improve the reliability of downstream gene prediction, SV detection, and PAV inference. Haplotype purging was not performed before graph construction; therefore, polished primary assemblies were used directly for pan-genome graph construction.

Assembly contiguity statistics including assembled size, contig count, and contig N50, were calculated from the final polished assemblies. Gene-space completeness was assessed with BUSCO v5.7.1 [56] using the embryophyta_odb10 dataset, with eukaryota_odb10 included as an additional check. Complete BUSCOs were reported as the primary metric. Assemblies were considered suitable for graph-based pan-genome construction when they showed high gene-space completeness, broadly comparable contiguity among accession groups, and no extreme deviations in assembly size or repeat content. Per-accession sequencing depth, ONT read N50, assembly size, contig number, contig N50, BUSCO completeness, and repeat-content estimates are reported in Supplementary Table S2, while group-wise statistical comparisons and quality-related covariate analyses are provided in Supplementary Table S3.

Repeat annotation

To characterize repeat landscapes and minimize confounding effects in downstream analyses, repeat libraries were built de novo using RepeatModeler v2.0.7 [57]. Repeat masking was then performed with RepeatMasker v4.2.0 [58] and the custom library, producing masked assemblies and per-sample repeat-content estimates. These estimates were used to evaluate repeat-content variation among assemblies and to assess whether assembly contiguity or downstream structural variation patterns could be influenced by repeat-rich regions. Statistical analyses evaluating the relationship between repeat content, sequencing metrics, and assembly contiguity are summarized together with assembly-quality group comparisons in Supplementary Table S3.

Graph-based pan-genome construction

A graph-based pan-genome was constructed from the 30 polished oil palm assemblies using the cactus-pangenome workflow from the Comparative Genomics Toolkit v2.9.5 [14], with EG11 [8] specified as the primary reference backbone. Each accession assembly was provided as an individual haploid assembly path, and assemblies were integrated by the Minigraph-Cactus strategy, which first builds an initial pangenome graph from whole-genome alignments and then refines local alignments to represent alternative haplotypes and structural variants. EG11 was used as the coordinate backbone for graph construction and all coordinate-based downstream analyses.

The graph was generated in GBZ/GFA-compatible formats and converted to standard VG formats using the vg toolkit v1.58.0 [19]. Graph summary statistics, including graph size, node statistics, graph component metrics, and coverage properties, are provided in Supplementary Table S4. To characterize graph-level expansion, node-growth rarefaction curves were generated with panacus v0.4.1 [59] using incremental genome additions and repeated random permutations.

The graph was used as the basis for downstream structural and gene-content analyses. Graph-derived variant records were exported in VCF format and normalized with bcftools norm for SV classification, per-accession SV burden estimation, chromosome-level SV density analysis, gene-overlap analysis, PAV inference, and insertion-derived CNV gain analysis. Gene coordinates, SV localization, CNV origin assignment, and functional annotation of reference genes were all interpreted relative to the EG11 reference coordinate system. The complete graph construction and conversion commands are provided in Supplementary Note S2.

Structural variant definition and normalization

To standardize variant representation across accessions, multi-allelic sites in the graph-derived VCF were decomposed into biallelic records using bcftools norm (-m -any). Because the graph-derived SV VCF lacked explicit SVLEN and SVTYPE annotations, variant length was defined as $L = \max(\text{len}(\text{REF}), \text{len}(\text{ALT}))$, with SV50 operationally defined as variants where $L \geq 50$ bp. Each variant was assigned a unique identifier ($\text{VAR_KEY} = \text{CHROM}:\text{POS}:\text{REF}:\text{ALT}$), which was consistently applied to join genotype tables, BED intervals, and sequence extractions.

Insertions and deletions were classified directly from allele-length differences in the normalized SV50 records. Insertions were defined as records where $\text{len}(\text{ALT}) > \text{len}(\text{REF})$, whereas deletions were defined as records where $\text{len}(\text{REF}) > \text{len}(\text{ALT})$, retaining only events with allele length ≥ 50 bp. Duplication-associated variants were inferred as a subset of insertion events whose inserted sequences showed strong alignment to the EG11 reference genome, indicating that the inserted sequence likely originated from another genomic locus. Insertion sequences were aligned to EG11 using minimap2 v2.30 (-x asm5), and only alignments with ≥ 0.90 sequence identity and ≥ 0.80 coverage were retained as high-confidence duplication-derived candidates.

Duplication-associated insertions were then classified according to the spatial relationship between the insertion site and its inferred origin locus. Tandem duplications were defined as events where the insertion site occurred within or near the inferred origin locus on the same

chromosome, including events within 10 kb. Proximal duplications were defined as same-chromosome events located >10 kb and ≤ 200 kb from the inferred origin. Dispersed duplications were defined as same-chromosome events located >200 kb from the inferred origin. Interchromosomal duplications were defined as events where the insertion site and inferred origin were located on different chromosomes or chromosome-level sequences. These duplication categories follow the general convention of classifying duplicated sequences by physical proximity to their inferred origin, adapted here to graph-derived insertion-origin mappings rather than whole-gene family phylogenies. They were therefore treated as operational spatial classes for describing inferred duplication-associated structural variation, not as direct evidence of specific molecular mechanisms. Because interchromosomal classes depend on origin assignments across different chromosome-level sequences, they were considered particularly sensitive to assembly fragmentation, repetitive-sequence alignment ambiguity, retained haplotigs, misjoins, or graph construction artifacts, and were therefore interpreted with additional caution.

Per-accession SV burden was summarized by counting non-reference genotypes for each accession from the normalized graph-derived SV50 records. Per-accession SV counts by variant class are provided in Supplementary Table S5.

Chromosome-level SV density was calculated to account for differences in chromosome length. Insertion and deletion events were counted per EG11 chromosome from graph-derived SV50 records, whereas duplication-associated events were assigned to chromosomes based on insertion cluster coordinates. For each chromosome, SV density was calculated as the number of events divided by chromosome length in megabases. Raw chromosome-level SV counts and normalized densities are provided in Supplementary Table S6.

Gene coordinate harmonization for gene-level SV mapping

Gene coordinates were obtained from the EG11 reference annotation (GFF3) and converted to BED format using gene features (type = gene). Coordinates were represented as 0-based half-open intervals and retained gene identifiers for downstream overlap analyses. SV50 intervals were similarly formatted as BED records, with start = POS - 1 and end = start + L.

Gene presence-absence variation matrix and pan-genome categories

Gene presence-absence variation was inferred using EG11 reference gene models as the primary coordinate framework rather than by independently predicting complete gene models for each assembly. This approach was chosen to ensure consistent gene identifiers, coordinates, and category assignments across all accessions, thereby enabling direct comparison of gene occupancy, rarefaction, functional enrichment, and R-gene distribution. Reference-gene PAV was therefore interpreted as variation affecting EG11-annotated genes. To partially capture gene content absent from EG11, insertion-derived candidate non-reference protein-coding loci were analyzed separately using a conservative ORF, homology, and transposable element filtering workflow described below.

To quantify gene-level presence-absence patterns across accessions, SV50 intervals were intersected with EG11 gene coordinates using bedtools intersect, generating mappings between gene IDs and overlapping variant keys. Genotypes for each SV50 were extracted from the

normalized VCF with bcftools query and recoded into a ternary scheme: 1 (alternate allele observed), 0 (reference homozygous), and NA (missing/no-call).

Gene presence-absence inference was derived from SVs overlapping EG11 gene bodies. Deletion-associated SVs contributed directly to absence calls, whereas insertion-derived SVs were used to represent allele-specific gene disruption or gain depending on their overlap and genotype status. Missing genotypes were treated conservatively according to variant class: missing genotypes for deletion-derived variants were treated as absence, whereas missing genotypes for insertion-derived variants were treated as presence to avoid spurious loss inference due to graph incompleteness or alignment ambiguity. Inference was performed using a custom Python workflow (`make_gene_matrix.py`), with the option `--all-na-as-present true` applied only to insertion-derived variants.

Genes were classified into pan-genome categories using occupancy thresholds: Core - present in ≥ 0.95 of samples, Shell - present in ≥ 2 samples but below the core threshold, and Unique - present in exactly one sample, including reference-specific genes present only in EG11 after filtering. The resulting binary PAV matrix and category assignments were used for clustering, rarefaction analysis, and functional comparisons. Gene counts and percentages for core, shell, and unique categories are provided in Supplementary Table S7.

Gene-level expansion was assessed from randomized rarefaction curves of the binary PAV matrix by evaluating the accumulation of total, core, shell, and unique genes across repeated genome-order permutations. Because graph node growth reflects structural expansion and does not directly measure gene-level pan-genome openness, this analysis was interpreted separately from graph node growth to avoid conflating structural graph expansion with gene-level pan-genome expansion. Heaps' law modelling was then used to quantify whether total gene discovery continued at the current sampling depth.

Assembly-based validation of gene presence-absence variation

To independently validate gene presence-absence inference, genic sequences containing exon-intron structure were extracted from EG11 [8] reference annotations and used as queries. Alignments against each genome assembly were performed with *minimap2* v2.30 [60], and gene presence was defined when query sequences aligned with $\geq 85\%$ coverage and $\geq 95\%$ identity. These thresholds ensured high-confidence matches while minimizing spurious alignments to paralogous regions or fragmented repeats. Validation performance was summarized using confusion-matrix metrics, with interpretation focused on false-absence rates rather than overall accuracy, which is sensitive to class imbalance. Category-specific validation metrics, including confusion matrix statistics for all genes and for core, shell, and unique categories, are reported in Supplementary Table S7.

Copy number gain inference from insertion-derived duplications

Copy-number gains associated with insertion variants were inferred from SV50 records where $\text{len(ALT)} > \text{len(REF)}$ and $\text{len(ALT)} \geq 50$ bp. Insertion sequences were aligned to the EG11 reference genome using *minimap2* v2.30 [60] (`-x asm5`) to identify likely origin loci. Alignments were filtered to retain strong duplication candidates with ≥ 0.90 identity and ≥ 0.80 coverage.

For each retained insertion, carrier samples were identified from genotype tables based on non-reference genotypes. Insertion origin mappings were combined with carrier lists to generate a

gene-by-sample copy-number gain matrix. For downstream comparative analysis, copy-number gain was encoded as binary variable indicating where a given accession carried an insertion-derived duplication associated with a given gene. This binary representation was used to summarize gain incidence and was not interpreted as absolute copy-number dosage.

Copy-number gain summaries were generated per accession, domestication group, and duplication class. Duplication classes followed the spatial criteria defined above for tandem, proximal, dispersed, and interchromosomal duplication-associated insertions. Accession-level CNV burden, group-level gain proportions, and gene-level CNV frequencies are provided in Supplementary Table S8.

Insertion-derived candidate non-reference protein-coding loci

To characterize non-reference gene content contributed by insertion variants, we focused on sequences lacking confident genomic origin alignments to the EG11 reference genome (“no-origin” insertions). Because plant genomes are rich in transposable elements (TEs), candidate loci were identified using a conservative, multi-evidence strategy designed to prioritize gene-like coding potential while excluding TE-derived artifacts.

No-origin insertion sequences were normalized to compact identifiers and used to predict open reading frames (ORFs), generating protein sequences with insertion- and ORF-level identifiers. Homology screening of all predicted ORFs was performed against the SwissProt protein database using stringent similarity criteria, retaining only strong matches based on strict e-value and coverage thresholds to minimize short, spurious, or low-quality alignments.

To explicitly exclude transposable element (TE)-derived sequences, ORFs were screened using complementary evidence sources. First, protein sequences were scanned against the Pfam-A database with *HMMER* to identify TE-associated domains, including reverse transcriptase (RVT), RNase H, integrase, DDE transposases, hAT/MULE, Helitron-related domains, and Gag/zf-CCHC motifs. Second, ORFs were flagged as TE-like based on conservative keyword rules applied to homology annotations (e.g., “reverse transcriptase,” “integrase,” “retrotransposon,” “LINE/LTR,” “transposase,” “Helitron”). Combination logic was applied to avoid excluding unrelated cellular enzymes (e.g., RNase H in isolation). ORFs supported by any TE-consistent evidence were removed.

Because individual insertion loci may yield multiple ORFs, homology screening and TE exclusion were performed at the ORF level as an evidence-filtering step rather than as a direct gene-counting procedure. The remaining non-TE ORFs were integrated at the level of insertion-derived candidate loci, which were designated as candidate non-reference protein-coding loci.

Presence-absence of candidate loci was inferred by projecting insertion carrier status from the SV genotype table onto the corresponding loci. Missing genotypes were treated as absence (NA = 0). The resulting matrix was merged with the reference gene presence-absence matrix to generate a unified gene-content PAV matrix for pan-genome categorization and downstream functional analyses. Candidate loci were interpreted conservatively relative to reference-annotated genes but retained as meaningful contributors to gene-content variation. Nevertheless, because the primary PAV framework was anchored to EG11 gene models, highly diverged or lineage-specific genes not represented in EG11 and not captured as high-confidence insertion-derived candidate loci may remain underrepresented.

Functional annotation and enrichment analysis

To investigate functional differentiation among pan-genome components, protein-coding genes were annotated using homology-based approaches. Predicted protein sequences from the reference genome and novel candidates were processed with *eggNOG-mapper* v2.1.13 [61] against the eggNOG database to assign orthology-based functional annotations, including Gene Ontology (GO) terms, eggNOG-derived COG functional categories, and KEGG pathway annotations where available. In parallel, KEGG Orthology (KO) annotation was performed with *KOfamScan* v1 [62] using profile hidden Markov models and adaptive score thresholds. KO assignments from both tools were integrated into a consensus KO set at the gene level, retaining only annotations that passed default quality filters.

COG categories were retained only as broad, eggNOG-derived functional classes to summarize high-level differences among pan-genome components, rather than as a standalone or primary annotation database. Fine-scale biological interpretation was based primarily on GO terms and KEGG/KO annotations. Genes were grouped into pan-genome categories (core, shell, unique) as defined by PAV analyses. Functional enrichment was evaluated by comparing the distribution of GO terms, eggNOG-derived COG categories, and KEGG pathways across these categories. For each functional term or pathway, observed counts in each pan-genome category were compared against genome-wide expectations using contingency-based enrichment tests. Raw p-values were adjusted for multiple testing using the Benjamini–Hochberg procedure, and terms with $FDR < 0.05$ were considered statistically significant. Full enrichment statistics, including observed and expected counts, enrichment ratios, raw p-values, and FDR values, are provided in Supplementary Table S9.

Resistance gene annotation and classification

Resistance genes (R-genes) were identified using *DRAGO-API*, which detects plant disease resistance genes based on conserved domain architectures. Protein sequences were screened for characteristic resistance-related domains, including nucleotide-binding site (NBS), leucine-rich repeat (LRR), receptor-like kinase (RLK), receptor-like protein (RLP), and mildew locus O (MLO). Identified R-genes were classified into major classes according to domain composition, including nucleotide-binding leucine-rich repeat (NLR)-related classes and receptor-like classes. CNL genes were defined as coiled-coil NLRs and therefore represent a subclass of NLR genes, whereas TNL genes were defined as Toll/interleukin-1 receptor-like NLRs and NL genes as NLR-like genes lacking clearly assigned CC or TIR domains. Additional classes included receptor-like kinases (RLKs), receptor-like proteins (RLPs), and mildew locus O (MLO) genes.

R-gene annotations were integrated with pan-genome presence–absence and copy-number variation data to examine patterns of stability and variability across accessions. For each accession, total R-gene repertoire size, class-level R-gene counts, and the number of R-genes affected by copy-number gain were quantified. Group-level comparisons among wild, semi-domesticated, and commercial accessions were performed to assess differences in R-gene abundance and CNV burden. Accession-level R-gene repertoires, class-level counts, pan-genome category assignments, and R-gene CNV gain burden are provided in Supplementary Table S10.

Pan-genome category and CNV integration for R-genes

To evaluate the distribution of R-gene classes across pan-genome components, each annotated R-gene was assigned to a category (core, shell, or unique) based on its presence–absence profile. Copy-number information was integrated to identify R-genes exhibiting gains in one or more accessions.

Associations between R-gene classes and pan-genome categories were assessed using contingency analyses. Enrichment or depletion of specific classes within pan-genome components was evaluated from standardized residuals comparing observed versus expected counts. These results were used to position R-gene classes along a stability–variability spectrum, contrasting conserved receptor-mediated defense components with more dynamic intracellular resistance genes. Enrichment and depletion statistics for R-gene classes across pan-genome categories, including observed counts, expected counts, standardized residuals, raw p-values, and FDR values, are reported in Supplementary Table S11. Descriptive stability–variability profiles, including the proportion of core, shell, and unique genes for each R-gene class, are provided in Supplementary Table S12.

Statistical analysis and visualization

All statistical analyses were conducted in R v4.5.1 [63] unless otherwise stated. Group-level comparisons of assembly metrics, copy number variation burden, and R-gene counts were performed using non-parametric tests implemented in the stats package [63] (Kruskal-Wallis test) and the rstatix package [64] for post hoc pairwise comparison where applicable. Correlation analyses were conducted using Spearman’s rank correlation coefficient as implemented in stats package. Multiple testing correction was applied using the Benjamini–Hochberg procedure in the stats package. The complete statistical outputs supporting assembly-quality analyses, functional enrichment analyses, and R-gene enrichment tests are reported in Supplementary Tables S3, S9, and S11, respectively.

Data visualization was performed using ggplot2 [65], with additional plotting and layout refinements implemented using ggbeeswarm[68], cowplot[69], and patchwork[70]. All plots were generated directly from the processed data matrices used in downstream analyses, ensuring consistency between figures and reported statistics.

Terminology notes

To ensure reproducibility and minimize ambiguity, all technical terms in this study are applied consistently as defined in the Supplementary Note S1. For example, copy number variation (CNV) is restricted to structural variant–derived copy number gains, pan genome components are classified as core, shell, or unique genes, and resistance gene classes are abbreviated after initial definition. Readers are referred to the Supplementary Note S1 for complete definitions and standardized usage across the manuscript.

Author contributions

R.A., H.A.S., and S.S. conceived and designed the study and coordinated the project. S.S. and C.U. supervised the research and provided overall project guidance. R.A. generated genome assemblies and performed structural variant analyses. E.S., H.M.R., and Z.A.T. contributed to data interpretation and manuscript review. T.L., D.D., S.W.A., W.B.S., and H.M.R. served on

the supervisory committee and contributed to data interpretation and manuscript review. All authors approved the final version of the manuscript.

Data Availability

Genome assemblies generated in this study have been submitted to NCBI GenBank under BioProject accession PRJNA1394696 and are scheduled for public release upon publication. Individual BioSample and GenBank assembly accession numbers, together with release status for each oil palm accession, are provided in Supplementary Table S1. The accession-level sequencing and assembly statistics supporting the pan-genome construction are provided in Supplementary Table S2. Additional processed data supporting the findings of this study, including graph-derived variant summaries, gene presence–absence matrices, copy-number gain summaries, functional enrichment outputs, and resistance gene analyses, are provided in the Supplementary Tables or are available from the corresponding author upon reasonable request.

Ethics, Consent to Participate, and Consent to Publish declarations

Not applicable.

Acknowledgment

We appreciate the management of Bogor Botanical Gardens and the National Research and Innovation Agency (BRIN) for giving access to the Dura-BBG sample. We acknowledge PT SMART Tbk for the complementary financial support and supplementary resources essential for the study implementation. We also thank the Indonesian Oil Palm Research Institute (IOPRI) for providing oil palm accessions and facilitating this study. AI was used to polish the English language quality in the manuscript preparation.

Competing interests

All authors declare no competing interests.

Funding

- Redi Aditama, The Oil Palm Research Grant, the Indonesian Plantation Fund Management Agency (BPDP), GRS_20220227110350
- Heri Adriwan Siregar, The Oil Palm Research Grant, the Indonesian Plantation Fund Management Agency (BPDP), GRS_20220227110350
- Diny Dinarti, The Oil Palm Research Grant, the Indonesian Plantation Fund Management Agency (BPDP), GRS_20220227110350
- Sudarsono Sudarsono, The Oil Palm Research Grant, the Indonesian Plantation Fund Management Agency (BPDP), GRS_20220227110350

References

1. Murphy DJ. Agronomy and Environmental Sustainability of the Four Major Global Vegetable Oil Crops: Oil Palm, Soybean, Rapeseed, and Sunflower. *Agronomy*. 2025;15:1465. <https://doi.org/10.3390/agronomy15061465>
2. Jamshidi EJ, Yusup Y, Hooy CW, Kamaruddin MA, Mat Hassan H, Muhammad SA, et al. Predicting oil palm yield using a comprehensive agronomy dataset and 17 machine learning and deep learning models. *Ecol Inform*. 2024;81:102595. <https://doi.org/10.1016/j.ecoinf.2024.102595>
3. Lieke S-D, Spiller A, Busch G. Can consumers understand that there is more to palm oil than deforestation? *Sustain Prod Consum*. 2023;39:495–505. <https://doi.org/10.1016/j.spc.2023.05.037>
4. Gaveau DLA, Locatelli B, Salim MA, Husnayaen, Manurung T, Descals A, et al. Slowing deforestation in Indonesia follows declining oil palm expansion and lower oil prices. Zang R, editor. *PLOS ONE*. 2022;17:e0266178. <https://doi.org/10.1371/journal.pone.0266178>
5. Zakaria L. Basal Stem Rot of Oil Palm: The Pathogen, Disease Incidence, and Control Methods. *Plant Dis*. 2023;107:603–15. <https://doi.org/10.1094/PDIS-02-22-0358-FE>
6. Flood J, Bridge PD, Pilotti CA. Basal stem rot of oil palm revisited. *Ann Appl Biol*. 2022;181:160–81. <https://doi.org/10.1111/aab.12772>
7. Ong A-L, Teh C-K, Mayes S, Massawe F, Appleton DR, Kulaveerasingam H. An Improved Oil Palm Genome Assembly as a Valuable Resource for Crop Improvement and Comparative Genomics in the Arecoideae Subfamily. *Plants*. 2020;9:1476. <https://doi.org/10.3390/plants9111476>
8. Low E-TL, Chan K-L, Zaki NM, Taranenko E, Ordway JM, Wischmeyer C, et al. Chromosome-scale *Elaeis guineensis* and *E. oleifera* assemblies: comparative genomics of oil palm and other Arecaceae. Jamann T, editor. *G3 Genes Genomes Genet*. 2024;14:jkae135. <https://doi.org/10.1093/g3journal/jkae135>
9. Singh R, Low E-TL, Ooi LC-L, Ong-Abdullah M, Ting N-C, Nagappan J, et al. The oil palm SHELL gene controls oil yield and encodes a homologue of SEEDSTICK. *Nature*. 2013;500:340–4. <https://doi.org/10.1038/nature12356>
10. Della Coletta R, Qiu Y, Ou S, Hufford MB, Hirsch CN. How the pan-genome is changing crop genomics and improvement. *Genome Biol*. 2021;22:3. <https://doi.org/10.1186/s13059-020-02224-8>
11. Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat Commun*. 2016;7:13390. <https://doi.org/10.1038/ncomms13390>
12. Matthews CA, Watson-Haigh NS, Burton RA, Sheppard AE. A gentle introduction to pangenomics. *Brief Bioinform*. 2024;25:bbae588. <https://doi.org/10.1093/bib/bbae588>
13. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. *Nature*. 2023;617:312–24. <https://doi.org/10.1038/s41586-023-05896-x>

- 1099 14. Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph
1100 construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol.* 2024;42:663–
1101 73. <https://doi.org/10.1038/s41587-023-01793-w>
- 1102 15. Groza C, Schwendinger-Schreck C, Cheung WA, Farrow EG, Thiffault I, Lake J, et al.
1103 Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nat*
1104 *Commun.* 2024;15:657. <https://doi.org/10.1038/s41467-024-44980-2>
- 1105 16. Li X, Wang Y, Cai C, Ji J, Han F, Zhang L, et al. Large-scale gene expression alterations
1106 introduced by structural variation drive morphotype diversification in *Brassica oleracea*. *Nat*
1107 *Genet.* 2024;56:517–29. <https://doi.org/10.1038/s41588-024-01655-4>
- 1108 17. Li W, Liu J, Zhang H, Liu Z, Wang Y, Xing L, et al. Plant pan-genomics: recent advances,
1109 new challenges, and roads ahead. *J Genet Genomics.* 2022;49:833–46.
1110 <https://doi.org/10.1016/j.jgg.2022.06.004>
- 1111 18. Shi J, Tian Z, Lai J, Huang X. Plant pan-genomics and its applications. *Mol Plant.*
1112 2023;16:168–86. <https://doi.org/10.1016/j.molp.2022.12.009>
- 1113 19. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph
1114 toolkit improves read mapping by representing genetic variation in the reference. *Nat*
1115 *Biotechnol.* 2018;36:875–9. <https://doi.org/10.1038/nbt.4227>
- 1116 20. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, et al. Pangenomics
1117 enables genotyping of known structural variants in 5202 diverse genomes. *Science.*
1118 2021;374:abg8871. <https://doi.org/10.1126/science.abg8871>
- 1119 21. Du Z-Z, He J-B, Jiao W-B. A comprehensive benchmark of graph-based genetic variant
1120 genotyping algorithms on plant genomes for creating an accurate ensemble pipeline. *Genome*
1121 *Biol.* 2024;25:91. <https://doi.org/10.1186/s13059-024-03239-1>
- 1122 22. Grytten I, Rand KD, Nederbragt AJ, Sandve GK. Assessing graph-based read mappers
1123 against a baseline approach highlights strengths and weaknesses of current methods. *BMC*
1124 *Genomics.* 2020;21:282. <https://doi.org/10.1186/s12864-020-6685-y>
- 1125 23. Mayes S, Jack PL, Corley RHV. The use of molecular markers to investigate the genetic
1126 structure of an oil palm breeding programme. *Heredity.* 2000;85:288–93.
1127 <https://doi.org/10.1046/j.1365-2540.2000.00758.x>
- 1128 24. Barcelos E, Rios SDA, Cunha RNV, Lopes R, Motoike SY, Babiychuk E, et al. Oil palm
1129 natural diversity and the potential for yield improvement. *Front Plant Sci* [Internet]. 2015 [cited
1130 2025 Oct 3];6. <https://doi.org/10.3389/fpls.2015.00190>
- 1131 25. Gan ST, Teo CJ, Manirasa S, Wong WC, Wong CK. Assessment of genetic diversity and
1132 population structure of oil palm (*Elaeis guineensis* Jacq.) field genebank: A step towards
1133 molecular-assisted germplasm conservation. Avramidou EV, editor. *PLOS ONE.*
1134 2021;16:e0255418. <https://doi.org/10.1371/journal.pone.0255418>
- 1135 26. John Martin JJ, Yarra R, Wei L, Cao H. Oil Palm Breeding in the Modern Era: Challenges
1136 and Opportunities. *Plants.* 2022;11:1395. <https://doi.org/10.3390/plants11111395>

1137 27. Hayati A, Wickneswari R, Maizura I, Rajanaidu N. Genetic diversity of oil palm (*Elaeis*
1138 *guineensis* Jacq.) germplasm collections from Africa: implications for improvement and
1139 conservation of genetic resources. *Theor Appl Genet.* 2004;108:1274–84.
1140 <https://doi.org/10.1007/s00122-003-1545-0>

1141 28. MacNish TR, Al-Mamun HA, Bayer PE, McPhan C, Fernandez CGT, Upadhyaya SR, et
1142 al. *Brassica* Panache: A multi-species graph pangenome representing presence absence
1143 variation across forty-one *Brassica* genomes. *Plant Genome.* 2025;18:e20535.
1144 <https://doi.org/10.1002/tpg2.20535>

1145 29. Morgante M, Depaoli E, Radovic S. Transposable elements and the plant pan-genomes.
1146 *Curr Opin Plant Biol.* 2007;10:149–55. <https://doi.org/10.1016/j.pbi.2007.02.001>

1147 30. Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new
1148 reference. *Nat Plants.* 2020;6:914–20. <https://doi.org/10.1038/s41477-020-0733-0>

1149 31. Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan CK, et al. The pangenome
1150 of hexaploid bread wheat. *Plant J.* 2017;90:1007–13. <https://doi.org/10.1111/tpj.13515>

1151 32. Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, et al. Insights
1152 into the Maize Pan-Genome and Pan-Transcriptome. *Plant Cell.* 2014;26:121–35.
1153 <https://doi.org/10.1105/tpc.113.119982>

1154 33. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, et al. Pan-genome analysis highlights the
1155 extent of genomic variation in cultivated and wild rice. *Nat Genet.* 2018;50:278–84.
1156 <https://doi.org/10.1038/s41588-018-0041-z>

1157 34. Song J-M, Guan Z, Hu J, Guo C, Yang Z, Wang S, et al. Eight high-quality genomes reveal
1158 pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat Plants.* 2020;6:34–
1159 45. <https://doi.org/10.1038/s41477-019-0577-7>

1160 35. Henson IE. A Brief History of the Oil Palm. *Palm Oil* [Internet]. Elsevier; 2012 [cited 2023
1161 Aug 14]. p. 1–29. <https://doi.org/10.1016/B978-0-9818936-9-3.50004-6>

1162 36. Purba AR, Noyer JL, Baudouin L, Perrier X, Hamon S, Lagoda PJJ. A new aspect of
1163 genetic diversity of Indonesian oil palm (*Elaeis guineensis* Jacq.) revealed by isoenzyme and
1164 AFLP markers and its consequences for breeding. *Theor Appl Genet.* 2000;101:956–61.
1165 <https://doi.org/10.1007/s001220051567>

1166 37. Corley RHV, Tinker PB. *The Oil Palm*. 5th ed. Wiley-Blackwell; 2016.

1167 38. Cochard B, Adon B, Rekima S, Billotte N, De Chenon RD, Koutou A, et al. Geographic
1168 and genetic structure of African oil palm diversity suggests new approaches to breeding. *Tree*
1169 *Genet Genomes.* 2009;5:493–504. <https://doi.org/10.1007/s11295-009-0203-3>

1170 39. Yu P, Wang C-H, Xu Q, Feng Y, Yuan X-P, Yu H-Y, et al. Genome-wide copy number
1171 variations in *Oryza sativa* L. *BMC Genomics.* 2013;14:649. [https://doi.org/10.1186/1471-](https://doi.org/10.1186/1471-2164-14-649)
1172 2164-14-649

1173 40. Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, et al. Pervasive
1174 gene content variation and copy number variation in maize and its undomesticated progenitor.
1175 *Genome Res.* 2010;20:1689–99. <https://doi.org/10.1101/gr.109165.110>

- 1176 41. Springer NM, Anderson SN, Andorf CM, Ahern KR, Bai F, Barad O, et al. The maize W22
1177 genome provides a foundation for functional genomics and transposon biology. *Nat Genet.*
1178 2018;50:1282–8. <https://doi.org/10.1038/s41588-018-0158-0>
- 1179 42. Li Y, Zhou G, Ma J, Jiang W, Jin L, Zhang Z, et al. De novo assembly of soybean wild
1180 relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol.*
1181 2014;32:1045–52. <https://doi.org/10.1038/nbt.2979>
- 1182 43. Tettelin H, Massignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome
1183 analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the
1184 microbial “pan-genome.” *Proc Natl Acad Sci.* 2005;102:13950–5.
1185 <https://doi.org/10.1073/pnas.0506758102>
- 1186 44. Shivute FN, Zhong Y, Wu J, Bao Y, Wang W, Liu X, et al. Genome-wide and pan-genomic
1187 analysis reveals rich variants of NBS-LRR genes in a newly developed wild rice line from
1188 *Oryza alta* Swallen. *Front Plant Sci.* 2024;15:1345708.
1189 <https://doi.org/10.3389/fpls.2024.1345708>
- 1190 45. Van De Weyer A-L, Monteiro F, Furzer OJ, Nishimura MT, Cevik V, Witek K, et al. A
1191 Species-Wide Inventory of NLR Genes and Alleles in *Arabidopsis thaliana*. *Cell.*
1192 2019;178:1260-1272.e14. <https://doi.org/10.1016/j.cell.2019.07.038>
- 1193 46. Silva-Arias GA, Gagnon E, Hembrom S, Fastner A, Khan MR, Stam R, et al. Patterns of
1194 presence–absence variation of NLRs across populations of *Solanum chilense* are clade-
1195 dependent and mainly shaped by past demographic history. *New Phytol.* 2025;245:1718–32.
1196 <https://doi.org/10.1111/nph.20293>
- 1197 47. Jones JDG, Vance RE, Dangl JL. Intracellular innate immune surveillance devices in plants
1198 and animals. *Science.* 2016;354:aaf6395. <https://doi.org/10.1126/science.aaf6395>
- 1199 48. Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, et al. Graph pangenome captures missing
1200 heritability and empowers tomato breeding. *Nature.* 2022;606:527–34.
1201 <https://doi.org/10.1038/s41586-022-04808-9>
- 1202 49. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication
1203 using fastp. *iMeta.* 2023;2:e107. <https://doi.org/10.1002/imt2.107>
- 1204 50. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform.
1205 *Bioinformatics.* 2009;25:1754–60. <https://doi.org/10.1093/bioinformatics/btp324>
- 1206 51. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of
1207 SAMtools and BCFtools. *GigaScience.* 2021;10:giab008.
1208 <https://doi.org/10.1093/gigascience/giab008>
- 1209 52. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation
1210 PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:s13742-015-
1211 0047–8. <https://doi.org/10.1186/s13742-015-0047-8>
- 1212 53. Cook DE, Andersen EC. VCF-kit: assorted utilities for the variant call format. Stegle O,
1213 editor. *Bioinformatics.* 2017;33:1581–2. <https://doi.org/10.1093/bioinformatics/btx011>

1214 54. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary
1215 Genetics Analysis across Computing Platforms. Battistuzzi FU, editor. *Mol Biol Evol.*
1216 2018;35:1547–9. <https://doi.org/10.1093/molbev/msy096>

1217 55. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic
1218 tree display and annotation tool. *Nucleic Acids Res.* 2024;52:W78–82.
1219 <https://doi.org/10.1093/nar/gkae268>

1220 56. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using
1221 repeat graphs. *Nat Biotechnol.* 2019;37:540–6. <https://doi.org/10.1038/s41587-019-0072-8>

1222 57. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An
1223 Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly
1224 Improvement. Wang J, editor. *PLoS ONE.* 2014;9:e112963.
1225 <https://doi.org/10.1371/journal.pone.0112963>

1226 58. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:
1227 assessing genome assembly and annotation completeness with single-copy orthologs.
1228 *Bioinformatics.* 2015;31:3210–2. <https://doi.org/10.1093/bioinformatics/btv351>

1229 59. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2
1230 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci.*
1231 2020;117:9451–7. <https://doi.org/10.1073/pnas.1921046117>

1232 60. Tarailo-Graovac M, Chen N. Using RepeatMasker to Identify Repetitive Elements in
1233 Genomic Sequences. *Curr Protoc Bioinforma* [Internet]. 2009 [cited 2025 Feb 19];25.
1234 <https://doi.org/10.1002/0471250953.bi0410s25>

1235 61. Parmigiani L, Garrison E, Stoye J, Marschall T, Doerr D. Panacus: fast and exact
1236 pangenome growth and core size estimation. Kelso J, editor. *Bioinformatics.* 2024;40:btac720.
1237 <https://doi.org/10.1093/bioinformatics/btac720>

1238 62. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor.
1239 *Bioinformatics.* 2018;34:3094–100. <https://doi.org/10.1093/bioinformatics/bty191>

1240 63. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-
1241 mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the
1242 Metagenomic Scale. Tamura K, editor. *Mol Biol Evol.* 2021;38:5825–9.
1243 <https://doi.org/10.1093/molbev/msab293>

1244 64. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al.
1245 KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score
1246 threshold. Valencia A, editor. *Bioinformatics.* 2020;36:2251–2.
1247 <https://doi.org/10.1093/bioinformatics/btz859>

1248 65. R Core Team. R: A Language and Environment for Statistical Computing [Internet].
1249 Vienna, Austria: R Foundation for Statistical Computing; 2025. <https://www.R-project.org/>

1250 66. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests [Internet].
1251 2025. <https://cran.r-project.org/web/packages/rstatix/>

- 1252 67. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New
1253 York; 2016. <https://ggplot2.tidyverse.org>
- 1254 68. Clarke E, Sherrill-Mix S, Dawson C. ggbeeswarm: Categorical Scatter (Violin Point) Plots
1255 [Internet]. 2025. <https://github.com/eclarke/ggbeeswarm>
- 1256 69. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for “ggplot2” [Internet].
1257 2025. <https://wilkelab.org/cowplot/>
- 1258 70. Pedersen TL. patchwork: The Composer of Plots [Internet]. 2025.
1259 <https://github.com/thomasp85/patchwork>
- 1260

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [pg30xsupplementaryrevv1clean.docx](#)
- [pg30xsupplementarytblrevv1.xlsx](#)