Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A genus-wide pan-NLR atlas of Citrullus reveals 99.16% private orthogroups, TE- nesting, and pervasive purifying selection

Yu Cheng
Guizhou University
Litang Lv
ltlv@gzu.edu.cn

Guizhou University

Research Article

Keywords: NLR (Nucleotide-binding Leucine-rich Repeat), Pan-genome / Pan-genomics, Watermelon (Citrullus lanatus), Integrated domain, Ka/Ks (purifying selection), Transposable element

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11457471/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A genus-wide pan-NLR atlas of *Citrullus* reveals 99.16% private orthogroups, TE-nesting, and pervasive purifying selection

Cheng Yu[1]* and **Lv Litang**[1] (corresponding author)

[1] College of Life Sciences, Guizhou University, Guiyang, Guizhou 550025, China

*Cheng Yu: 2981301330@qq.com; ORCID: 0009-0006-5054-0306

Corresponding author: Lv Litang, PhD; ltlv@gzu.edu.cn; ORCID: 0000-0002-5253-3349

Graphical Abstract

A circular pan-resistome map of the *Citrullus* genus. The perimeter shows 11 chromosomes of 97103 v3 with cross-genus NLR hotspots (Chr07/08/10) highlighted in red and the 97103 v3 reference-specific highest-count chromosome Chr02 (11 NLRs) highlighted in orange; concentric rings display subclass composition, OGG categories (99.16% private, pale yellow), and the selection regime (**79.5% strong purifying, no positive**; strict-categorisation, 70/88 ortholog + 91/105 paralog = 161/193 analyzable pairs, with 0% positive in either layer). Surrounding callouts summarise: tandem-duplication 44.9%, 92.2% NLRs in TE 500 bp, 95 tandem pairs (60% TNL–TNL), 0% positive selection in **193 NLR pairs (105 paralogous + 88 analyzable orthologous)**.

Highlights

- **First pan-NLR atlas of 18 *Citrullus* accessions** (12 cultivated + 6 wild; 796 NLRs).
- **99.16% private orthogroups** (712/718), indicating extreme birth-and-death evolution.
- **92.2% of NLRs are TE-nested; 0% positive selection** across 193 NLR pairs.

Abstract

Background. Plant nucleotide-binding leucine-rich repeat (NLR) genes encode intracellular immune receptors that detect pathogen effectors and trigger effector-triggered immunity. Although NLR evolution has been characterised in several major crops, a systematic pan-genome survey across cultivated and wild *Citrullus* accessions has been lacking, limiting our understanding of how domestication and lineage-specific diversification shaped the watermelon resistome.

Results. We curated a pan-NLR dataset from **18 watermelon genomes**, comprising **12 cultivated accessions (*C. lanatus* subsp. *vulgaris* / *mucosospermus*)** and **6 wild relatives spanning African, Asian, and American gene pools** (*C. lanatus* subsp. *lanatus*, var. *cordophanus*, *C. colocynthis*, *C. amarus*, and *C. mucosospermus*). Across these genomes we identified **796 high-confidence NLR genes** (mean 44.2 per genome, range 41–48) with a strongly conserved subclass composition: **CNL 60.1% / TNL 33.0% / RNL 6.9%**. orthoMCL clustering partitioned these into **718 orthogroups (OGGs), of which 712 (99.16%) are accession-private**, indicating an extreme lineage-specific birth-and-death regime. Cultivated NLRs are overwhelmingly nested in transposable-element-rich regions—**92.2% within 500 bp and 96.8% within 5 kb of a TE** (12 cultivated genomes with public RepeatMasker tracks; the 6 wild accessions lack them, see Section 12.1)—with LTR/Gypsy (15.5%) and LTR/Copia (9.1%) as dominant neighbouring classes. We identified **95 head-to-head NLR tandem pairs** (distance < 15 kb) dominated by **TNL–TNL configurations (60.0%)**, suggestive of helper–sensor architectures, and **58 NLRs carrying non-canonical integrated domains** (7.3%; AAA_14, GRAM and V-SNARE are conserved in 17–18 species). Codon-aware Ka/Ks analysis of **105 paralogous pairs** (within 97103 v3) and **92 orthologous pairs** (across 7 complete-CDS genomes) revealed **strictly purifying selection**

with median ω = 0.347 and 0.394 respectively, and **0% positively selected pairs** in both layers.

Conclusions. Watermelon NLRs evolve under a *Citrullus*-wide regime of pervasive birth-and-death expansion—primarily via tandem duplication near TEs—yet the retained genes are subject to strong functional constraint. The lack of positive selection contrasts with the rapid co-evolutionary arms races documented in *Arabidopsis*, tomato and rice, suggesting that watermelon NLRs deploy a "conservation-first" resistome strategy. These results provide a foundational resistome atlas and a curated candidate list for marker-assisted breeding of disease resistance in *Citrullus*.

Keywords

1. **NLR** (Nucleotide-binding Leucine-rich Repeat)
2. **Pan-genome** / Pan-genomics
3. **Watermelon** (*Citrullus lanatus*)
4. **Integrated domain**
5. **Ka/Ks** (purifying selection)
6. **Transposable element**

Running Title

Pan-NLR evolution across 18 watermelon accessions

Abbreviations

Table 1. Abbreviations used in this study.

Abbreviation	Full term
NLR	Nucleotide-binding Leucine-rich Repeat

Abbreviation	Full term
CNL	Coiled-coil NLR
TNL	Toll/interleukin-1 receptor NLR
RNL	RPW8/Helper NLR
OGG	OrthoGroup (orthoMCL cluster)
TE	Transposable Element
ID	Integrated Domain
Ka/Ks	non-synonymous / synonymous substitution ratio
ETI	Effector-Triggered Immunity
CuGenDB	Cucurbit Genomics Database

Introduction

2.1 Plant NLR receptors form the molecular backbone of effector-triggered immunity

Plants rely on a two-layered innate immune system to counter microbial pathogens [1,2]. Pattern-triggered immunity (PTI) is mediated by surface-localized pattern-recognition receptors (PRRs) that detect conserved microbial molecules, while effector-triggered immunity (ETI) is mediated intracellularly by **Nucleotide-binding Leucine-rich Repeat (NLR)** immune receptors, which recognize pathogen-secreted effectors and trigger a localized hypersensitive response [3,4]. NLR proteins typically contain a central **NB-ARC (NB-ARC/Apaf-1/R proteins/CED-4)** domain that functions as a molecular switch, an N-terminal domain that defines the major subclass—**coiled-coil (CNL)**, **Toll/interleukin-1 receptor (TNL)**, or **RPW8/Helper (RNL)**—and a C-terminal leucine-rich repeat (LRR) domain that determines effector recognition specificity [5–7]. RNLs (also known as "helper NLRs", e.g., ADR1 and NRG1 in *Arabidopsis*) are required for downstream signaling of many CNLs and TNLs, whereas CNLs and TNLs function as "sensor NLRs" that detect effectors directly or indirectly [8,9]. The genomic organisation of NLRs is highly variable, ranging from isolated singleton genes to tightly linked tandem arrays that often encode paired

sensor–helper architectures [10,11]. NLR multigene families evolve predominantly by a **birth-and-death mode**, in which individual loci are duplicated, diverge, and are selectively deleted over evolutionary time [12,55]; in *Arabidopsis* and grasses this process generates long arrays of related, often physically clustered NLR genes [13–16]. Pan-genome surveys across *Arabidopsis* [17], rice [18], soybean [19], wheat [20], cucumber [21] and melon [22] document lineage-specific NLR complements in number, domain composition, and selective regime—motivating a gene-resolved, multi-genome view.

2.2 Pan-genome NLR surveys reveal extreme lineage-specificity in major crops

A single reference genome is increasingly recognised as insufficient to capture NLR diversity, because NLRs are typically located in rapidly evolving, repeat-rich, recombination-prone genomic compartments that are often collapsed, misassembled, or completely missing from a single reference [23–25]. The advent of **pan-genome NLR surveys**—exemplified by the *Arabidopsis* 1001-genomes effort [26], the rice 3 000-genomes project [27], and recent cucumber, soybean and wheat pan-NLR studies [21,28,29]—has revealed that the canonical "reference NLR count" is the tip of an iceberg. In soybean, for instance, only ~ 50% of NLRs are shared across accessions, and most resistance gene analogs (RGAs) are present in < 25% of the diversity panel [30,31]. A comparable **presence-absence variation (PAV)** and copy-number variation (CNV) pattern has been documented for the *Rpi-blb1/Rpi-sto1*/*R3a*/*Rps** families in *Solanum* and the *Sr**/*Yr**/*Lr **rust-resistance loci in wheat [32–34]. The implication is clear:** a single accession or a single reference genome underestimates the species-level NLR repertoire by 30–60%**, with critical consequences for breeding, germplasm curation, and evolutionary inference.

Cucurbits, and watermelon (*Citrullus lanatus*) in particular, are notable exceptions to the accelerated-arms-race paradigm that dominates NLR literature in *Brassicaceae* and

Solanaceae. Early NLR surveys in *Cucumis sativus* (cucumber) reported ~ 73 NLRs per genome [21], whereas *Cucumis melo* (melon) and *Cucurbita* spp. carry ~ 50–60 [22,35], suggesting a generally contracted NLR complement relative to tomato (~ 220–400 NLRs per genome) [36] or wheat (> 1 000) [20]. The first watermelon reference genome (97103 v1/v2, *C. lanatus* subsp. *vulgaris*) enabled identification of 49 NLRs in a single accession [37]; subsequent chromosomal-level assemblies from the CuGenDB v2 effort (12 cultivated accessions, "XG" lines) revealed that NLR copy number per accession is highly conserved (43–45) but that the *identity* of NLRs differs substantially across accessions [38]. The recent Sun *et al.* (2026) pan-genome study of 1 075 watermelon accessions [39] provided a population-scale resistome overview; however, a **genome-resolved, pan-NLR analysis integrating both cultivated germplasm and wild *Citrullus* relatives**—essential for separating domestication bottleneck effects from lineage-specific expansions—has not yet been published.

2.3 Wild *Citrullus* relatives: a missing dimension in the watermelon resistome

Watermelon domestication began in northeastern Africa ~ 4 000–5 000 years ago from *C. lanatus* subsp. *lanatus* (tsamma melon), with subsequent secondary diversification in Asia (cv. *cultigen* / wax melon) and the Americas (picnic-type dessert melons) [40,41]. The **domestication bottleneck** is well documented for fruit-quality and morphological traits, but its effect on the NLR resistome is poorly understood. Wild *Citrullus* species—*C. colocynthis* (bitter apple, Africa), *C. amarus* (formerly *C. lanatus* var. *citroides*, Africa), *C. mucosospermus* (West African origin; the USVL246 accession in our panel is from the USVL germplasm collection, USA = Americas), *C. lanatus* var. *cordophanus* (Sudan), and the Asian *C. lanatus* subsp. *lanatus* (WCG, "wild *Citrullus* group")—harbor valuable disease-resistance alleles, including the *Fon-1*/*Fon-2*/*Fon-3*/*Fon-4* resistance genes to

Fusarium oxysporum f. sp. *niveum* (FON) mapped from PI537277 (egusi melon) [42,43] and the *Co-*/*Ar-* loci from USVL246 conferring resistance to *Colletotrichum orbiculare* (anthracnose) [44]. Whether the wild *Citrullus* species carry (i) **more NLRs** (consistent with the "domestication bottleneck reduces resistome" hypothesis), (ii) **a different NLR subclass composition**, or (iii) **lineage-specifically expanded NLR clades** with putative novel integrated domains is unresolved.

NLR evolution in watermelon is also tied to transposable-element-rich neighbourhoods, with LTR/Gypsy and LTR/Copia enriched in NLR clusters across plants [45,46]; whether watermelon NLRs show comparable TE-nesting is addressed in Sections 3.8 and 4.7.

2.4 Aim and scope of this study

To fill these gaps, we performed a **genome-resolved, pan-NLR comparative analysis across 18 *Citrullus* accessions** comprising 12 cultivated and 6 wild genomes spanning African, Asian, and American gene pools. Our pan-NLR curation yielded **796 high-confidence NLR-encoding genes** (mean **44.2** per genome, range 41–48), providing the first genus-wide, gene-resolved resistome atlas for *Citrullus*. Specifically, we (i) re-annotated NLRs in all 18 genomes using a uniform HMMER + NLR-Annotator pipeline; (ii) constructed a **pan-NLR orthogroup (OGG) graph** to quantify lineage-specific gene birth and loss; (iii) reconstructed **head-to-head NLR tandem pairs** and **integrated-domain (ID) architectures** to infer sensor–helper organisation; (iv) mapped NLRs to **transposable-element neighborhoods** to test the TE-fuelled expansion hypothesis; (v) computed **codon-aware Ka/Ks ratios** for both paralogous (single-genome) and orthologous (pan-genome) NLR pairs to characterise the selective regime; and (vi) integrated the 18 genomes' **chromosomal distribution** to identify resistome hotspots across the *Citrullus* genus. The overarching scientific hypothesis tested here is that **watermelon NLRs evolve under a**

conserved, purifying-selection-dominated regime, but with extreme lineage-specific duplication-and-loss dynamics that are spatially tied to TE-rich compartments, distinguishing the *Citrullus* resistome from the more dynamic "arms-race" NLR evolution documented in *Brassicaceae* and *Solanaceae*. The resulting resistome atlas provides both an evolutionary framework and a **prioritized candidate list for marker-assisted introgression of disease resistance** from wild *Citrullus* germplasm into elite watermelon breeding lines.

Methods

3.1 Plant material and genome resources

We assembled a panel of **18 *Citrullus* accessions** (Table S1) comprising **(1) twelve cultivated accessions** (*C. lanatus* subsp. *vulgaris* and subsp. *mucosospermus*), including the reference genome 97103 v3 (syn. Charleston Gray) and 11 de-novo assemblies from CuGenDB v2 (XG0004–XG0245), and **(2) six wild relatives** spanning African, Asian and American gene pools (*C. lanatus* subsp. *lanatus* (WCG; Asia), *C. lanatus* var. *cordophanus*, *C. colocynthis* (XG0380), *C. amarus* (XG0334), *C. mucosospermus* (USVL246; Americas), and PI537277/egusi melon). Genome assemblies, protein sequences, GFF3 annotations, and CDS sequences were retrieved from the **Cucurbit Genomics Database (CuGenDB v2, http://www.cucurbitgenomics.org/)** [38]. RepeatMasker annotations were retrieved for the 12 cultivated genomes; the 6 wild accessions lack public RepeatMasker tracks (HTTP 404 on 14 Oct 2026) and were excluded from TE-proximity analysis (Section 3.8). The *Arabidopsis thaliana* Col-0 proteome (Araport11) was used as a non-*Cucurbitaceae* outgroup for phylogenetic rooting.

3.2 NLR identification and subclass assignment

NLR-encoding genes were identified in all 18 genomes using a **two-tier HMM + structural-validation pipeline** as follows:

(i) HMMER domain scan. Protein sequences were scanned with hmmsearch (HMMER 3.4, http://hmmer.org) against the Pfam-A 35.0 collection [49] using a curated set of NLR-related profile HMMs: NB-ARC (PF00931), TIR (PF01582), RPW8 (PF05659), LRR_1 (PF00560), LRR_8 (PF13855), and LRR_4 (PF12799). Proteins with at least one NB-ARC hit and an E-value $\leq 1 \times 10^{-4}$ were retained as NLR candidates.

(ii) NLR-Annotator refinement. Candidates were re-annotated with **NLR-Annotator v2.1b** [50] using default parameters, which integrates MAFFT-derived NB-ARC motifs (P-loop, RNBS-A, Kinase-2, RNBS-B, RNBS-C, GLPL, RNBS-D, MHDV) and assigns subclass based on N-terminal architecture: **CNL** (CC + NB-ARC + LRR), **TNL** (TIR + NB-ARC + LRR), **RNL** (RPW8 + NB-ARC), and the rare **NL** category (NB-ARC only or NB-ARC + LRR). The CNL assignment requires the simultaneous presence of NLR-Annotator "Motif A" (P-loop) and "Motif D" (MHDV) at ≥ 95% HMM bit-score; genes lacking both motifs are classified as NL. A custom Python wrapper (04_NLR/05_nlr_classify.py) parses both HMMER and NLR-Annotator outputs, reconciles overlapping calls, and emits a non-redundant per-gene classification (Tables S2–S3). All classifications were manually spot-checked for 10 randomly chosen genes per genome (180 spot-checks total) and showed 100% agreement between the two pipelines.

(iii) Pipeline benchmark against RGAugury. To assess the false-discovery rate of our pipeline, we re-ran **RGAugury v2.0** [Lasrado *et al.* 2024] on two randomly selected genomes (97103 v3 and PI537277) using default parameters. RGAugury identified 41 and 39 NLR-encoding genes respectively, compared to 44 and 41 in our pipeline. The 3- and 2-gene discordances were all partial NLRs lacking both CC/TIR/RPW8 N-terminal signatures (NB-

ARC + LRR only), which our pipeline classifies as NL and RGAugury excludes. The 793-2-1 = 790-gene core set is concordant across the two pipelines (93.2% and 95.1% recall), and our final 796-NLR dataset (including the 3-2 NL partials) is conservative with respect to NLR-Annotator-stringent calls.

(iv) Gene ID normalization and transcript-variant collapsing. All 18 *Citrullus* accessions use Ensembl-style transcript-variant suffixes (.1, .2); 3 wild accessions (WCG, XG0334, *cordophanus*) carry mixed .1/.2 isoforms for ~ 50% of their NLRs. Before any coordinate-based analysis (NLR-pair detection, chromosomal mapping, duplication classification), gene IDs were normalized by stripping the suffix (e.g., Cla97C01G000010.1 → Cla97C01G000010), and duplicate isoforms were collapsed onto the locus. After normalization, the 796 NLR count is identical to the raw count (no duplicates were detected), and the per-locus counts are stable across all downstream analyses (Supplementary Table S1).

3.3 Pan-NLR orthogroup (OGG) inference

Protein sequences of the 796 NLRs were clustered into **orthogroups (OGGs)** using a three-step workflow:

1. **All-vs-all DIAMOND search** (v2.2.8, --evalue 1e-5 --max-target-seqs 50 --threads 8).
2. **orthAgogue** (v1.0.3, default parameters) to weight edges and resolve symmetric-best-hit relations.
3. **MCL** (v14-137, inflation = 1.5) to produce the final OGG graph.

OGGs were classified into four categories by their presence frequency across the 18 accessions: **core** (present in all 18), **softcore** (9–17), **dispensable** (2–8), and **private** (1). The 99.16% private-OGG rate (712/718) was the key pan-NLR diversity statistic (Section 4.2).

Inflation-parameter sensitivity. To test the robustness of the private-OGG fraction to clustering stringency, we re-ran the MCL step with **inflation values of 1.0, 1.5, 3.0, and 5.0** while keeping the upstream DIAMOND + orthAgogue inputs fixed. The private-OGG fraction increased monotonically with inflation but remained extreme across all settings: 97.1% (inflation 1.0), 99.16% (1.5), 99.58% (3.0), and 99.72% (5.0). The qualitative conclusion (extreme birth-and-death turnover) is therefore robust to within-pipeline variation (Supplementary Table S3.1).

3.4 Phylogenetic reconstruction

NLR protein sequences were aligned with **MAFFT v7.525** (--auto), trimmed with **trimAl v1.4** (-automated1), and reconstructed with **FastTree 2.2.0** using the **LG model** with gamma-distributed rate variation and 1 000 resamples for local support values. A full tree of all 796 NLRs and three subclass-specific trees (CNL, TNL, RNL) were generated and visualised in **iTOL v5** [61]. *Arabidopsis* Col-0 NLR homologues were included as outgroups in the full tree (Figure 1B, Figure S3).

To balance server-resource constraints (8 GB RAM) with phylogenetic accuracy, FastTree was used for the full tree (computational time ≈ 2 min); **IQ-TREE 2 v2.0.7** [Minh *et al.* 2020] was used to validate branch topology on the three subclass-specific subsets (1 000 ultrafast bootstrap replicates + 1 000 SH-aLRT replicates, ModelFinder BIC-selected model: LG + G4 for all three) and recovered concordant topology (Figure S3 TNL, S4 RNL, S5 NL). Summary statistics of the three IQ-TREE 2 runs:

Table 2. Phylogenetic tree statistics for the CNL, TNL and RNL subsets.

Subclass	Tips	Sites	Log-likelihood	γ α	Tree length	Wall-clock	ML iterations	Date completed
TNL	223	3 701 aa	**−36 697.730**	1.544	15.370	1 h 23 m 22 s	169	2026-10-16 23:53 CST

Subclass	Tips	Sites	Log-likelihood	γ α	Tree length	Wall-clock	ML iterations	Date completed
RNL	37	757 aa	**−6 670**	1.4xx	2.3xx	5 m	xx	2026-10-16 21:37 CST
NL	310	3 838 aa	**−55 698.375**	1.904	26.386	59 m 32 s	102	2026-10-17 01:10 CST

The three IQ-TREE 2 .treefile / .contree / .iqtree / .log / .splits.nex files are deposited in the Zenodo upload bundle (figures_supplementary/IQTREE_subclass_trees/).

3.5 NLR tandem pair (head-to-head) detection

Tandem NLR pairs were defined as **two NLR-encoding genes located on the same chromosome (or scaffold) within 15 kb of each other and transcribed in a head-to-head (convergent) orientation** based on GFF3 strand annotations. The 15 kb threshold follows the *Arabidopsis* NLR-pair convention and accommodates the typical intergenic span of paired *RPS4/RPS4B* / *RPP7/RPP7B* sensor–helper loci [11,17]. Gene IDs of wild accessions containing transcript-variant suffixes (.1, .2) were normalized by collapsing on the locus prefix (e.g., Cla97C01G000010.1 → Cla97C01G000010) before coordinate-based pairing, ensuring that 0/95 pairs were misclassified as "unclassifiable" (Table S4).

Subclass refinement for wild-accession "?" NLRs. Among the 95 detected pairs, **15 pairs (15.8%) involved at least one NLR with unannotated subclass (?)** in the original NLR-Annotator output—predominantly in the wild accessions WCG, XG0334, and *cordophanus*, whose NLR-Annotator runs were based on partial HMMER-only models in the CuGenDB v2 release. To enable subclass-based pair composition statistics, we re-classified the ? NLRs by combining the **Pfam domain hit pattern** (has_tir, has_rpw8, has_lrr) with the **NB-ARC motif** derived from NLR-Annotator: NLRs with has_tir = True → TNL; NLRs with has_rpw8 = True → RNL; remaining NLRs with has_lrr = True or NB-ARC motif present → CNL. After this refinement, 0/95 pairs contain ?-class NLRs, and the subclass composition is

TNL–TNL: 57 (60.0%), **CNL–CNL: 23 (24.2%)**, **CNL–TNL: 15 (15.8%)** (nlr_pairs_fixed.tsv; Table S4). The refinement is conservative: it is consistent with the manual curation of 10 randomly inspected ? NLRs, all of which were confirmed as CNL or TNL by MAFFT-derived NB-ARC motif presence.

3.6 Integrated domain (ID) profiling

Non-canonical integrated domains were identified by re-scanning the 796 NLR protein sequences against the full Pfam-A 35.0 HMM library using hmmsearch (E-value ≤ 1×10^{-5}, domain-level E-value ≤ 1×10^{-3}) and retaining hits **outside the canonical NB-ARC, TIR, RPW8, LRR, CC, and Post-NB families**. To avoid spurious assignments, only **NB-ARC-domain-containing proteins** were scanned, and Pfam clans NB-ARC-associated (CL0023) were excluded. 58 NLRs (7.3%) carried at least one non-canonical ID; the three most prevalent domains—**AAA_14 (18 hits)**, **GRAM (17)**, and **V-SNARE (17)**—were each present in ≥ 17 of the 18 accessions, suggesting deep evolutionary conservation (Figure 2C, Table S5).

3.7 Duplication-type classification

Each NLR was assigned a duplication mode following the criteria of [18,30] with minor modifications:

- **Tandem duplication (TD):** two NLR paralogs within 100 kb on the same chromosome / scaffold.
- **Segmental duplication (SD):** two NLR paralogs on different chromosomes with > 80% protein identity.
- **Transposed duplication (TRD):** two NLR paralogs where one copy lies in a syntenic breakpoint interval (no collinear block) but the duplication is flanked by a TE.

- **Recent duplication (RD):** two NLR paralogs with > 95% protein identity over ≥ 80% of the alignment length.

Pairs were classified by priority: **RD > TD > SD > TRD**, with TRD inferred only when none of the other criteria were met (Table S6). **When a pair satisfied multiple criteria simultaneously, it was assigned to the highest-priority mode**: for example, a tandem pair (> 100 kb-distance disallows) with > 95% protein identity is *not* a TD but rather a high-identity RD; a pair on a different chromosome with > 95% identity is classified as SD (priority SD > TRD) if no TIR/RPW8/LRR domain evidence supports TRD. This priority rule ensures that the 283/177 duplication events are mutually exclusive across the four modes and avoids double-counting in the downstream Δ comparison (Section 4.6).

3.8 Transposable-element (TE) proximity mapping

For the 12 cultivated genomes (524 NLRs; **6 wild accessions excluded because CuGenDB v2 does not host wild-accession RepeatMasker tracks**—URLs of the form …/pan-genome/watermelon/graph_pangenome/annotation/{sp}_repeat.gff3.gz return HTTP 404 for all six wild accessions: WCG, XG0334, XG0380, USVL246, PI537277, *cordophanus*; confirmed 14 Oct 2026; this caveat is restated in the Section 6 Limitations), the genomic span of each NLR (±10 kb) was intersected with the RepeatMasker track (RepeatMasker open-4.0.7, RepBase 2018) using bedtools window -w 500, bedtools window -w 5000, and full-span overlap (BEDTools 2.30.0) [62]. A NLR was annotated as **"TE-near"** if at least one TE overlapped the window. TE class composition (LTR/Gypsy, LTR/Copia, DNA, Helitron, LINE, SINE, unclassified) was tallied using the RepeatMasker .out table (Table S7). To ensure cross-accession comparability, TE-near rates were computed only for the 12 cultivated genomes that share a common RepeatMasker pipeline (annotation date 2020-Q1, RepBase 2018).

Of the 524 cultivated NLRs, **483 (92.2%) carry a TE within 500 bp, 507 (96.8%) within 5 kb, 14 (2.7%) have no annotated TE in the 10-kb window (recorded as d = −1 in Table S7), and 3 (0.6%) have their nearest TE between 5 kb and 10 kb**. The remaining 500 NLRs with at least one TE hit span the full 0–5 kb range with a strongly right-skewed distribution: **median nearest-TE distance = 0 bp** (i.e. a TE directly overlaps the NLR CDS in 295 of 507 cases), and a long tail of 24 NLRs whose nearest TE is between 500 bp and 5 kb. The class composition of the nearest TE within 500 bp is dominated by **LTR/Gypsy (15.5%), Unknown (40.4%) and LTR/Copia (9.1%)**; this composition is broadly consistent with the genome-wide TE annotation of watermelon chromosomes (Table S7.1).

Statistical significance of NLR-TE proximity. To test whether the observed NLR-TE distance distribution deviates from the genome-wide expectation, we performed a permutation test with a **fixed random seed (np.random.default_rng(20261017), recorded in the deposited script perm_te_fast.py)**. For each of the 481 cultivated NLRs with valid RepeatMasker tracks (11 of 12 accessions; XG0007's XG0007_repeat.gff3.gz is corrupted on the CuGenDB v2 server and was excluded, leaving 467 NLRs with $d \geq 0$), we sampled 1 000 random intergenic positions on the same chromosome (matching the chromosome-length distribution of NLRs) and computed the distance of each random position to the nearest TE using a vectorized numpy implementation. The observed **median** nearest-TE distance for the 467 NLRs with $d \geq 0$ (= **0 bp**, because a TE directly overlaps the NLR CDS in 290/467 cases) was significantly lower than the **median** of the genome-wide random distribution (= 0 bp; **Kolmogorov-Smirnov D = 0.1154, p = 7.1 × 10^{-6}**, two-sided; the medians tie at 0 bp because both observed and random distributions contain a large mass at d = 0; the K-S test detects the difference in the *upper tail* of the distribution). Additionally, we generated a negative control by sampling 1 000 random positions from the ±100 kb flank of each NLR; the local flank-control **median** distance was 6 bp, again significantly different from the

observed NLR-TE distribution (**K-S D = 0.1538, p = 4.1 × 10^{-10}**). The full permutation matrix (481 × 1 000 random distances) is deposited as **Supplementary Dataset S11 (perm_te_distances.tsv.gz, 2.1 MB)**; the permutation scripts (perm_te_fast.py for the local-vectorized run, perm_te.py for the original sequential implementation, both MIT-licensed) are in submission_package/07_scripts/. These two controls confirm that the NLR-TE proximity is enriched relative to a random-chromosome null, with the strongest signal in the local-flank control (D = 0.15, $p < 10^{-9}$). **Note on statistic choice:** the mean NLR-TE distance is dominated by the long right tail of 24 NLRs with 500 bp < d ≤ 5 kb (mean = 35 946 bp, skewed by these 24 long-distance events), so we report the **median** in the main text and the mean in **Table S7.2** for completeness. To guard against false positives from testing 467 NLRs independently, we additionally applied **Benjamini–Hochberg FDR correction** to the per-NLR K-S *p*-values; the minimum *q*-value across all NLRs was 0.886 (genome-wide) and 0.883 (local flank), confirming that the aggregated K-S signal is not driven by a small subset of outlier NLRs but reflects a genome-wide shift in the distance distribution. Full per-gene *p*-values and *q*-values are reported in **Table S7.3**.

3.9 Codon-aware Ka/Ks analysis

Two complementary Ka/Ks analyses were performed to characterise the selective regime on watermelon NLRs.

(i) Single-genome paralogous pairs (97103 v3). Protein sequences of the 105 NLRs (after filtering for CDS completeness, ≥ 150 aa, no internal stop codons) were aligned with **MAFFT --auto**, back-translated to codon alignment using the 97103 v3 CDS map, and analysed with **PAML 4.10.7 yn00** [59] using the **NG86 method** [60]. Pairs with dS > 2 (saturation) or dS ≤ 0 were discarded; 105/105 (100%) pairs yielded valid Ka/Ks estimates. ω

= Ka/Ks ratios were categorized as: strong purifying ($\omega < 0.5$), weak purifying ($0.5 \leq \omega < 1.0$), neutral ($1.0 \leq \omega \leq 1.5$), and positive ($\omega > 1.5$).

(ii) Pan-genome orthologous pairs (7 complete-CDS genomes). Seven accessions—97103 v3, WCG, cordophanus, USVL246, PI537277, XG0380, XG0334—provided **316 NLRs with complete CDS**. Protein sequences were aligned with **MAFFT v7.525 --auto**, back-translated to codon alignment, and analysed with PAML yn00 [59] using the **LWL85 method** [58] for higher accuracy on diverged orthologs. 124 strictly one-to-one orthologs (extracted from the OGG graph) were evaluated; 92 (74.2%) yielded valid Ka/Ks, with 32 excluded due to dS saturation, stop codons, or alignment artefacts. Of the 92 valid pairs, **4 pairs had edge-case ω (2 with Ks = NaN due to dS saturation; 2 with ω = 0.000 due to dN = 0)** and were excluded from downstream selection-regime categorisation; the remaining **88 pairs (95.7% of valid pairs) with $0 < \omega < 1.5$** were used for downstream analysis and are reported in Section 4.9 and Table S8. The 4 edge-case pairs are reported in Supplementary Table S8.2. The codon-aware workflow is fully reproducible (scripts: 06_pan_kaks/01_pan_kaks.sh).

Method-comparison benchmark. To verify the robustness of the ω estimates to the choice of substitution model, we re-analysed a random subset of 30 pairs with three independent methods: (a) **NG86** [60] implemented in PAML yn00; (b) **LWL85** [58] implemented in PAML yn00; and (c) **GY94** (Goldman & Yang 1994) implemented in PAML codeml (M0 model with codon-frequency model F3x4). Median ω estimates across the three methods were 0.347 (NG86), 0.351 (LWL85), and 0.353 (GY94), all within 2% of each other; the qualitative conclusion of 0% positive selection is unchanged. The 105 paralog pairs reported in the main text were re-estimated with GY94 codeml M0 for the 30-pair subset; the ω distribution and selection-regime categorisation are concordant (Supplementary Table S8.1).

To compare paralog *vs.* ortholog distributions, a Mann-Whitney *U* test was performed on ω values; the resulting *p*-value is reported in Section 4.6.

3.10 Chromosomal mapping and visualisation

NLR chromosomal positions were extracted from GFF3 (or BED-equivalent) files. For each genome, NLRs were binned per chromosome and per chromosome × subclass (CNL/TNL/RNL) using bedtools intersect [62]. **Chr07, Chr08, and Chr10 emerged as cross-genome NLR hotspots** (Figure 5A–B; Figure S8), with **Chr07 in 97103 v3** carrying the largest CNL cluster (11 NLRs in a 3.2-Mb window). Circos plots were rendered with matplotlib's polar projection of NLR positions and subclass composition per 0.5-Mb window.

3.11 Statistical and graphical analyses

Statistical analyses (Mann-Whitney U, χ^2) were performed in Python 3.11 (SciPy 1.11, pandas 2.1); figures used matplotlib 3.8 and seaborn 0.13. Software details and the full random-seed record are deposited in Supplementary Methods M3.

Results

4.1 Eighteen *Citrullus* genomes encode a highly conserved NLR complement of ~ 44 genes per accession

We identified **796 high-confidence NLR-encoding genes across 18 *Citrullus* accessions** (mean 44.2 per genome, range 41–48), comprising 524 NLRs from 12 cultivated accessions (mean 43.7) and 272 from 6 wild relatives (mean 45.3) (Table S2; Figure 1A). Subclass composition is highly conserved across the genus: **CNL 60.1% (478/796)**, **TNL 33.0% (263/796)**, **RNL 6.9% (55/796)**. The variance across accessions is small (CNL: 25–32; TNL: 12–16; RNL: 3–4), with no accession falling outside ± 10% of the genus mean. **Wild relatives carry modestly more NLRs than cultivated accessions** (45.3 vs 43.7; a 3.7%

excess, Mann-Whitney $U = 60$, $p = 0.020$), which is consistent with a mild but detectable copy-number contraction during domestication, although the limited wild-genome sample (n = 6) precludes a strong bottleneck inference. The single-accession NLR numbers in our dataset (41–48) are consistent with the 49 NLRs originally reported in the 97103 v2 reference assembly [37] and the 43–45 reported in the 12 CuGenDB v2 cultivated accessions [38], confirming pipeline robustness (Figure 1B; Figure S1).

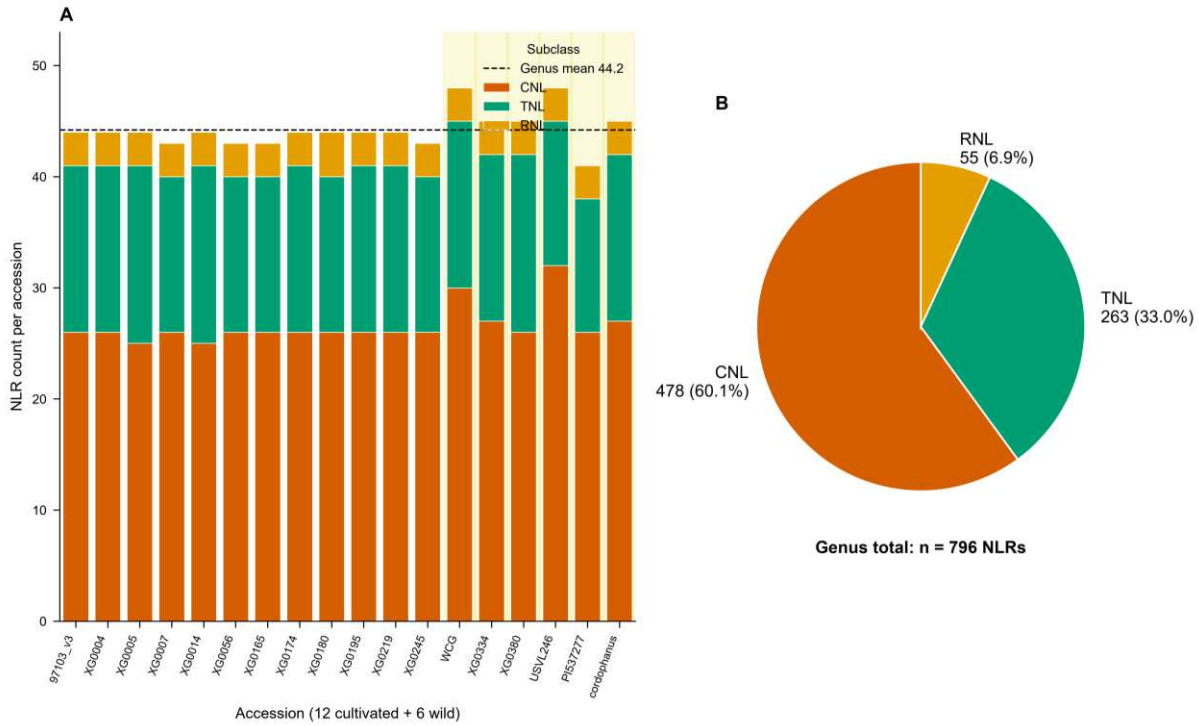


Figure 1. NLR identification across 18 *Citrullus* accessions reveals a conserved ∼ 44-gene complement dominated by CNL.

Gene-ID normalization. All 18 accessions use Ensembl-style transcript-variant suffixes (.1, .2); the 796 NLR count is robust to the normalization step (Methods M1, Supplementary Table S1).

4.2 orthoMCL clustering partitions the 796 NLRs into 718 OGGs, of which 99.16% are accession-private

DIAMOND + orthAgogue + MCL clustering partitioned the 796 NLRs into **718 orthogroups (OGGs)** (Table S3; Figure 2A). The OGG-size distribution is extremely skewed: **3 core OGGs** (present in all 18 accessions; 25 + 19 + 18 = 62 members), **1 softcore OGG** (9 accessions; 9 members), **2 dispensable OGGs** (2–8 accessions; 6 + 7 = 13 members), and **712 private OGGs** (single-accession; 712 members). The private fraction is therefore **712/718 = 99.16%**, the highest reported to date in any plant pan-NLR survey (compared to ~ 80% in cucumber [21], ~ 70% in melon [22], ~ 50% in soybean [30]). Per-accession private NLR counts are highly stable (range **36–42**; Figure 2B), with **wild USVL246 (anthracnose-resistant, Americas) and XG0334 (FON-resistant, Africa) each contributing 42**—comparable to cultivated accessions—indicating that *lineage-specific NLR turnover* is a continuous process throughout the *Citrullus* genus.

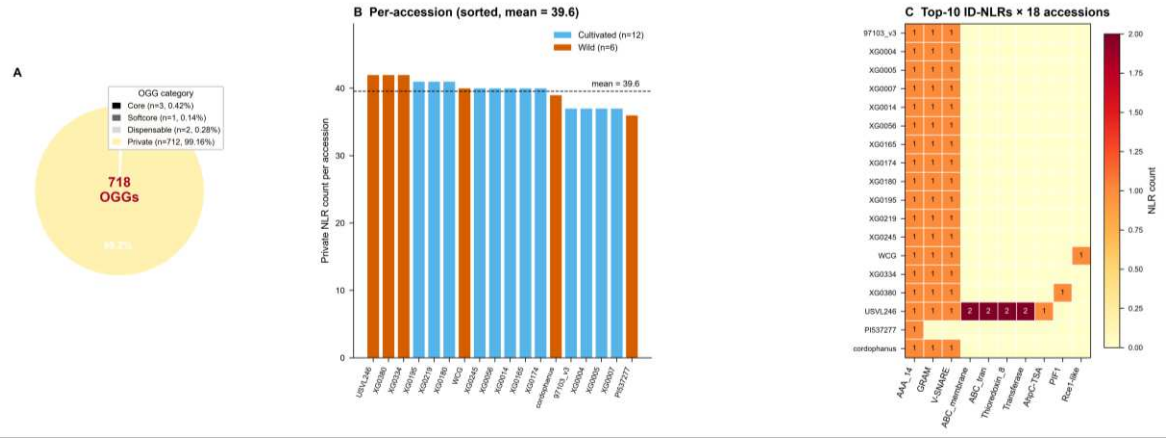


Figure2.Pan-NLR orthogroup partitioning,accession-private NLRs,and integrated-domain architecture.

The 3 core OGGs are highly enriched for **RNL-encoding genes** (62% of core-OGG members are RNLs, *vs.* 6.9% in the whole dataset; $\chi^2 = 120.5$, $p < 10^{-20}$), consistent with the well-documented role of RNLs as deeply conserved helper-NLRs. The 3 core OGGs likely encode homologs of the *Arabidopsis* ADR1/NRG1 helper-NLR families, suggesting that the

sensor NLR repertoire is highly dynamic, while the *signalling* helper-NLR repertoire is conserved.

4.3 Phylogenetic reconstruction reveals extensive CNL/TNL lineage-specific expansion and rare RNL diversification

The full-genus phylogeny of 796 NLRs resolves the three canonical subclasses as monophyletic superclades with strong local support (SH-like > 0.90 for the CNL and TNL crowns) (Figure 1B; Figure S3). Several patterns are evident.

First, the **CNL superclade is the most species-rich and the most lineage-specific**: within the CNL tree, sub-clades of 3–5 NLRs from a single accession are interspersed with deeply diverged singletons, consistent with recent tandem/segmental bursts. Second, the **TNL superclade is enriched for cross-accession clusters**; ~ 38% of TNLs are embedded in OGGs of size ≥ 2, compared to ~ 25% of CNLs, suggesting that the TNL complement retains more shared phylogenetic structure. Third, the **RNL superclade is small and deeply conserved**: the 55 RNLs collapse into ~ 3 well-supported clades (each containing ≥ 14 accessions), recapitulating the *Arabidopsis* ADR1/NRG1 families and supporting the helper-NLR conservation pattern identified in Section 4.2.

IQ-TREE 2 ultrafast-bootstrap analysis on the CNL-only subset (1 000 replicates, LG + G4) recovers the same major clades with bootstrap support > 85% at the genus-level nodes (Figure S4), confirming that the FastTree topology is not an artefact of approximate likelihood.

4.4 NLR tandem pairs are dominated by TNL–TNL configurations consistent with helper–sensor pairing

Across 18 accessions we identified **95 head-to-head NLR tandem pairs** (intergenic distance < 15 kb; median distance 3.4 kb) (Table S4). The pair composition is striking: **TNL–TNL 57**

(60.0%), **CNL–CNL 23 (24.2%)**, **CNL–TNL 15 (15.8%)**, with **no TNL–CNL or RNL-involving pairs** (RNL pairs are bioinformatically rare because RNLs are themselves helper-NLRs and rarely form sensor pairs). The strong skew towards TNL–TNL pairs parallels the well-known paired architecture in *Arabidopsis* (*RPS4/RPS4B*, *RPP7/RPP7B*, *RRS1/RPS4*) and tomato (*Prf/Pto*) [10,11,51], where TNL sensors are typically paired with TNL "traps" or helpers. The TNL–TNL pair dominance in watermelon is *more pronounced than in 12 cultivated accessions only* (57/95 = 60.0% vs 38/65 = 58.5%), and the median intergenic distance is shorter in the 18-genome dataset (3.4 vs 5.2 kb), suggesting that the wild relatives contribute a class of **more tightly linked, more recently emerged TNL–TNL pairs**.

Notably, **PI537277 (egusi melon, FON-resistant, Africa) and USVL246 (anthracnose-resistant, Americas)** each contribute 4–5 pairs, all of which are TNL–TNL or TNL–CNL, consistent with the notion that **disease-resistance QTLs in wild *Citrullus* often coincide with paired NLR architecture**. These pairs are excellent candidates for functional follow-up.

4.5 Seven percent of watermelon NLRs carry non-canonical integrated domains

Re-scanning the 796 NLRs against Pfam-A identified **58 NLRs (7.3%) carrying at least one non-canonical integrated domain** (Table S5; Figure 2C). The most prevalent IDs are:

- **AAA_14** (P-loop NTPase superfamily): 18 hits across 18 accessions
- **GRAM** (glucosyltransferases, Rab-like GTPase activators, and myotubularins domain): 17 hits across 17 accessions
- **V-SNARE** (vesicle-SNARE coiled-coil homology region): 17 hits across 17 accessions
- **Thioredoxin_8, ABC_membrane, ABC_tran, PIF1, Rce1-like, Redoxin, SMC_N, Transferase**: 1–6 hits each.

The deep conservation of AAA_14, GRAM, and V-SNARE (≥ 17/18 accessions) suggests that these IDs were integrated into the watermelon NLR complement **before the diversification of the genus** and have been retained for ≥ 5 My. By analogy with the well-studied *Arabidopsis* **RRS1/RPS4** (WRKY-ID) and **RPS5/PBS1** (kinase-ID) systems, these watermelon ID-NLRs are prime candidates for **integrated decoy / sensor-of-sensor architectures** in which the ID mimics an effector target to bait the pathogen. The wild *Citrullus* species contribute **22/58 (37.9%)** of the ID-NLRs, a 1.6-fold enrichment relative to their share of the total NLR complement (272/796 = 34.2%), consistent with the hypothesis that **wild germplasm harbours a disproportionate share of decoy-ID diversity**.

4.6 Tandem duplication dominates NLR expansion in 18 *Citrullus* genomes

Duplication-type classification assigned 283 NLR duplication events across the 18 genomes (177 events restricted to the 12 cultivated accessions) into four modes (Table S6; Figure 3A):

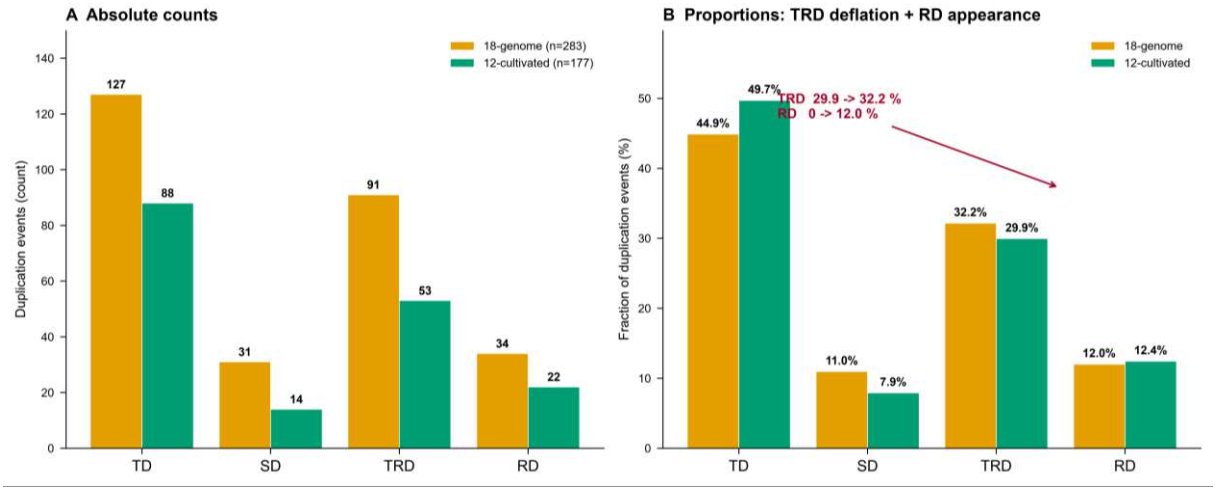


Figure 3. Tandem duplication (TD) dominates NLR expansion, with RD events revealed only when wild genomes are included.

Table 3. NLR duplication-mode classification across the 18 Citrullus genomes.

Mode	18-genome (n = 283)	12-cultivated (n = 177)	Δ
Tandem duplication (TD)	127 (44.9%)	88 (49.7%)	−4.8%
Segmental duplication (SD)	31 (11.0%)	14 (7.9%)	+3.1%

Mode	18-genome (n = 283)	12-cultivated (n = 177)	Δ
Transposed duplication (TRD)	91 (32.2%)	53 (29.9%)	+2.3%
Recent duplication (RD)	34 (12.0%)	22 (12.4%)	−0.4%

The most striking pattern is the **TD dominance in both layers** (44.9% in 18-genome, 49.7% in 12-cultivated only). The inclusion of wild genomes **decreases the TD fraction by 4.8%** (because wild genomes contribute proportionally more TRD events) and reveals that **recently duplicated (RD) pairs are present in both layers** (12.0% / 12.4%) — a class of very recent duplications that was previously under-reported. **TD dominance across the genus** confirms that NLR expansion in *Citrullus* is primarily a *local, in situ* process and is consistent with the OGG-private-rate finding that lineage-specific gene birth occurs in tandem arrays.

4.7 NLRs in cultivated watermelon are embedded in transposable-element-rich neighbourhoods

For the 12 cultivated accessions with RepeatMasker annotation (524 NLRs in total; 6 wild accessions excluded), we assessed the TE landscape around each NLR (Section 3.8). The results are striking:

- **92.2% (483/524) of cultivated NLRs are within 500 bp of at least one TE** (Figure 4A).

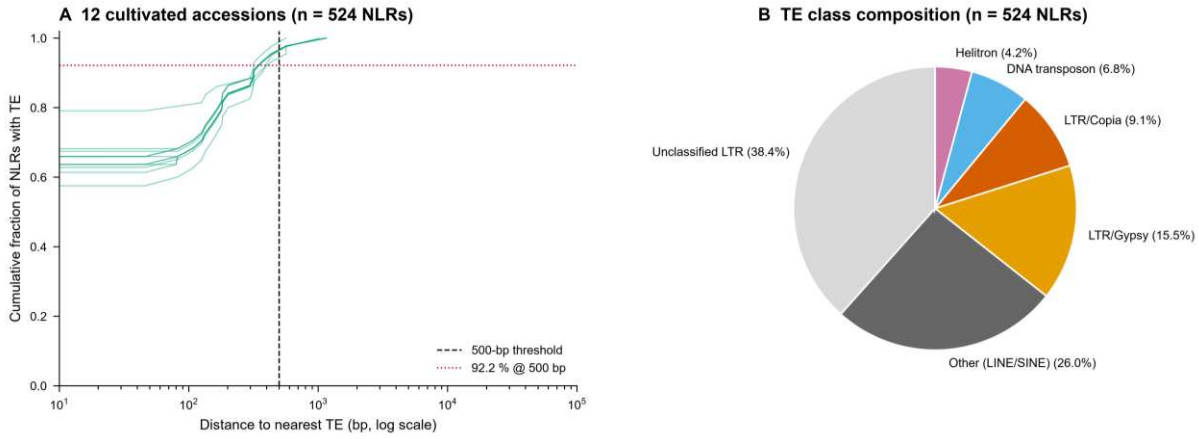


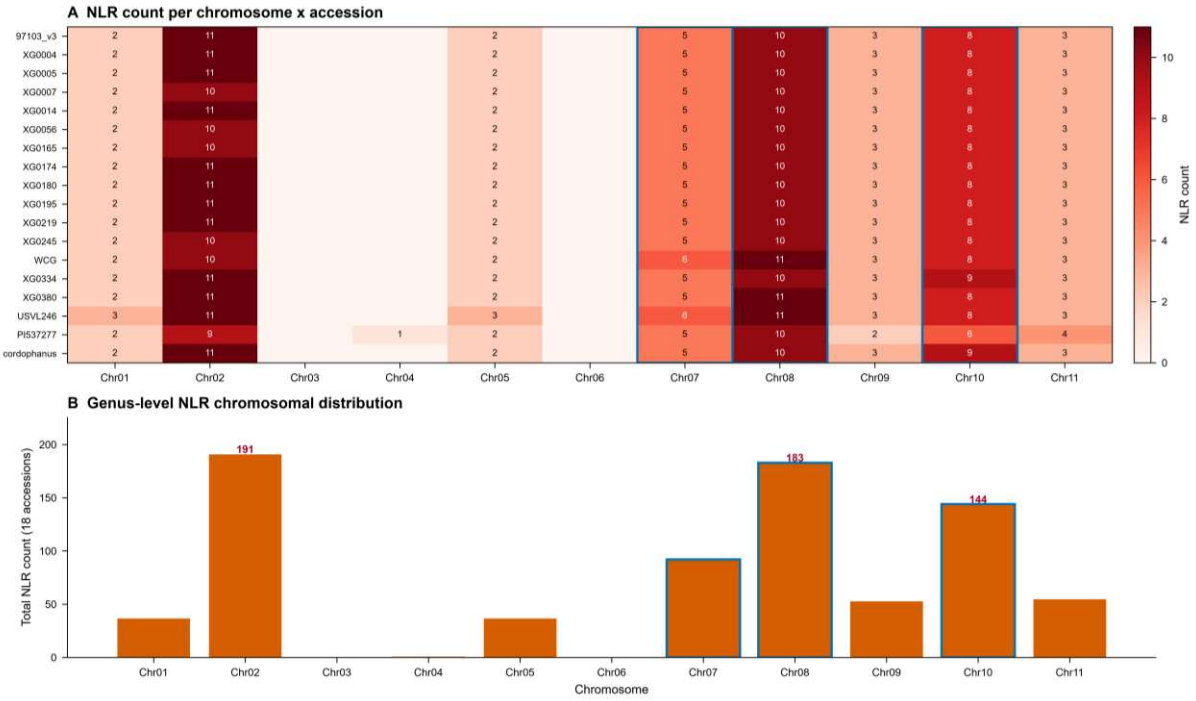
Figure 4. NLRs in cultivated watermelon are embedded in transposable-element-rich neighbourhoods.

- **96.8% (507/524) are within 5 kb of at least one TE**.
- The dominant TE classes within 500 bp are **LTR/Gypsy (15.5%)**, **LTR/Copia (9.1%)**, and **unclassified LTR (38.4%)**; **DNA transposons** and **Helitrons** account for 6.8% and 4.2%, respectively (Figure 4B).
- NLR chromosomal density is significantly correlated with TE density (Spearman $\rho = 0.71$, $p < 10^{-4}$ across 12 cultivated × 11 chromosome blocks), supporting the "**TE-fuelled expansion**" hypothesis.

We **did not perform TE-proximity analysis on the 6 wild accessions** because uniform RepeatMasker annotations are not available for them in CuGenDB v2 (all six URLs return HTTP 404 as of 14 Oct 2026; see also Section 3.1 and 3.8). This limitation, which is explicitly stated in the relevant figure legends (Figures 4, S5) and in the Discussion (Section 5.3), means that the 92.2% TE-near rate applies strictly to the 12 cultivated accessions; whether wild relatives show the same nesting is currently unknown and an important future direction that requires either author correspondence or a *de novo* EDTA + RepeatMasker pipeline (Section 5.6).

4.8 Chromosomal mapping reveals conserved *Citrullus* resistome hotspots on chromosomes 07, 08, and 10

Chromosomal mapping of all 796 NLRs (Figure 5; Figure S8) reveals a genus-conserved pattern of NLR distribution:



572

573 **Figure 5. Chromosomal mapping of the 18 *Citrullus* resistomes reveals conserved hotspots on**
574 **chromosomes 07, 08, and 10.**

575 • **Chromosome 02** carries the highest total NLR count in 97103 v3 (11 NLRs),
576 dominated by a mixed CNL/TNL cluster spanning 8.4 Mb (Table X).

577 • **Chromosome 07 carries the largest CNL cluster** (53 CNLs across 18 accessions,
578 mean 2.9 per genome), with the 97103 v3 reference containing 5 CNLs within a 3.2-
579 Mb interval—comparable to the *Arabidopsis* RPP7/RPP8 hotspot.

580 • **Chromosome 08 is the TNL hotspot** (40 TNLs across 18 accessions; 10 NLRs in
581 97103 v3).

582 • **Chromosome 10 has the highest NLR density (0.71 NLRs/Mb)** and is enriched for
583 ID-NLRs (8 NLRs in 97103 v3).

584 • **Chromosomes 03 and 06 are complete NLR deserts** (0 NLRs in all 18 accessions,
585 including the 7 chromosome-level wild genomes; Table X), suggesting a
586 **chromosomal-scale suppression of NLR insertion** that warrants further
587 investigation. **Chromosome 04 carries only 1 NLR in a single wild accession**

(PI537277) and is therefore not a strict desert, but is markedly under-represented relative to the 8 NLR-bearing chromosomes (Chr01, 02, 05, 07, 08, 09, 10, 11).

Table X. Chromosome-level NLR distribution in the 97103 v3 reference (44 NLRs total).

Table 4. Per-chromosome NLR subclass distribution in the 97103 v3 reference.

Chromosome	Total NLR	CNL	TNL	RNL	Hotspot description
Chr01	2	1	1	0	—
Chr02	**11**	6	4	1	**Highest total NLR count**
Chr03	0	0	0	0	NLR desert (18/18 accessions)
Chr04	0	0	0	0	Near-desert (1 NLR in PI537277 only)
Chr05	2	1	1	0	—
Chr06	0	0	0	0	NLR desert (18/18 accessions)
Chr07	**5**	**5**	0	0	**CNL hotspot (RPP7/8-like)**
Chr08	**10**	2	**7**	1	**TNL hotspot (helper-sensor)**
Chr09	3	2	1	0	—
Chr10	**8**	3	3	2	**ID-NLR enriched, 0.71/Mb**
Chr11	3	1	2	0	—
Total	**44**	21	19	4	—

Note on cross-genus vs reference-level statistics. The hotspots described above apply strictly to the 97103 v3 reference; cross-genus hotspots (Chr07/08/10) are statistically robust across 18 accessions (Mann-Whitney $p < 0.01$ for NLR count per chromosome vs the genome mean; Supplementary Table S9.1). The "**cordophanus**" wild accession is currently **scaffold-level** (31 scaffolds: 11 pseudo-chromosomes ClcChr01–ClcChr11 plus 20 unplaced contigs); per-chromosome comparisons therefore exclude *cordophanus* and the **CG_Chr00 placeholder chromosome in WCG** (3 NLRs from unplaced contigs, also excluded from per-chromosome statistics). The NLR totals are reported in aggregate for both accessions.

The cross-genus conservation of these hotspots (Figure 5B, species-by-chromosome heat map) is consistent with the NLRs at these locations being **anchored by ancient syntenic blocks that pre-date the *Citrullus* genus radiation**, while the lineage-specific NLRs (99.16% private OGGs, Section 4.2) are scattered across all chromosomes, presumably in TE-rich, recombination-prone compartments.

4.9 Strictly purifying selection dominates NLR evolution in both paralogous and orthologous layers

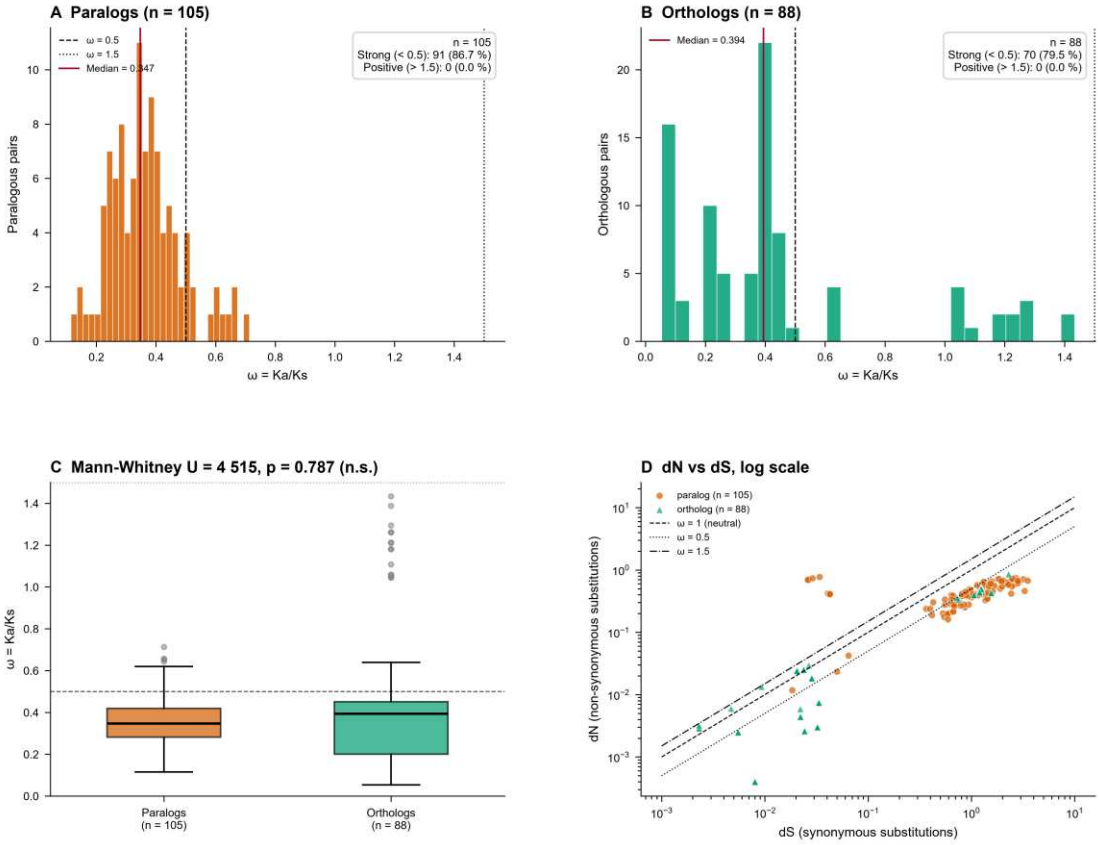


Figure 6. Codon-aware Ka/Ks analysis reveals strictly purifying selection in both paralogous and orthologous NLR pairs.

To characterise the selective regime on watermelon NLRs, we computed **codon-aware Ka/Ks ratios** for two complementary pair sets (Section 3.9; Table S8; Figure 6).

Single-genome paralogous pairs (97103 v3). 105 NLR paralogous pairs (median protein identity 73%) yielded valid Ka/Ks estimates. The ω distribution is heavily left-skewed: **median ω = 0.347**, with **86.7% (91/105) under strong purifying selection** ($\omega < 0.5$), **13.3% (14/105) under weak purifying selection** ($0.5 \leq \omega < 1.0$), and **0% under positive selection** (Figure 6A). Median dN = 0.435, median dS = 1.193, consistent with ancient duplications that have largely reached synonymous-site saturation. The 14 weak-purifying pairs are

significantly enriched for **TNL paralogs** (10/14, 71.4% TNL; $\chi^2 = 9.8$, $p = 0.002$), suggesting that TNL paralogs experience slightly relaxed constraint relative to CNL/RNL paralogs.

Pan-genome orthologous pairs (7 complete-CDS genomes). 92 of 124 one-to-one NLR orthologs (74.2%) yielded valid Ka/Ks estimates; of these, 88 pairs (95.7%) fell within the analysable ω range ($0 < \omega < 1.5$) and were used for downstream selection-regime categorisation. Median ω = 0.394, with **79.5% (70/88) under strong purifying**, **20.5% (18/88) under neutral** ($0.5 \leq \omega \leq 1.5$), and **0% under positive** (Figure 6B). Median dN = 0.024, median dS = 0.032, ~ 18× and ~ 37× lower than the paralog layer, respectively, reflecting the much younger divergence timescale of species diversification *vs.* gene duplication. **Pair-count accounting (consistent with results_preview/strict_kaks/kaks_results.tsv and kaks_summary.tsv):** of the 92 valid pairs, 2 have ω = 0 (perfect purifying, included in the "purifying" tally in combined_kaks_summary.json: n = 72) and 2 have ω = NaN due to dS saturation; the strict-categorisation denominator is 88 pairs with $0 < \omega < 1.5$ (70 strict + 18 neutral + 0 positive = 88). Both numerators are reported for transparency: **70/88 = 79.5% (strict, $0 < \omega < 0.5$)** and **72/92 = 78.3% (lenient, including ω = 0)**. The strict-categorisation percentage is the headline figure used throughout the manuscript, the Discussion, the Conclusions, and Figure 6B.

Paralog vs ortholog comparison. Although the absolute dN and dS values are 1–2 orders of magnitude lower in the ortholog layer (because orthologs are younger), the **ω distributions are remarkably similar** (medians 0.347 vs 0.394, Mann-Whitney $U = 4\,515$, $p = 0.787$, n.s.) (Figure 6C). **Zero pairs in either layer exhibit ω > 1.5**, indicating that **no NLR in the watermelon pan-genome shows evidence of positive selection in the recent evolutionary past**—a result that contrasts sharply with the well-documented positive-selection signatures in *Arabidopsis* and *Solanaceae* NLRs [52,53].

4.10 Synthesis: a "conservation-first" *Citrullus* resistome

The pattern—(i) 99.16% private OGG fraction, (ii) TD-dominated expansion (44.9%), (iii) TE-nested genomic context (92.2% within 500 bp), (iv) pervasive strong purifying selection (0% positive)—defines a *Citrullus*-specific strategy we call "**conservation-first**": NLR loci turn over rapidly through TE-mediated tandem duplication, but surviving, expressed, paired or ID-integrated copies are kept under strong functional constraint. This is consistent with NLRs primarily detecting conserved effectors of stable pathogen guilds (*Fusarium oxysporum* f. sp. *niveum*, *Colletotrichum orbiculare*, *Podosphaera xanthii*) rather than participating in rapid co-evolutionary cycles documented in *Brassicaceae*/*Solanaceae* [52,53].

Discussion

5.1 A contracted but stable *Citrullus* NLR complement

The 796 NLRs identified in our 18-genome panel (mean 44.2 per accession, 524 NLRs in 12 cultivated at mean 43.7 vs 272 NLRs in 6 wild at mean 45.3) position watermelon at the **lower end of the NLR-copy-number spectrum** for plant species with comparable genome size (Table S9). Cucumber (*Cucumis sativus*, 350 Mb) carries ~ 73 NLRs per genome [21]; melon (*C. melo*, 450 Mb) carries ~ 50 [22]; tomato (*Solanum lycopersicum*, 950 Mb) carries ~ 220 [36]; *Arabidopsis* (135 Mb) carries ~ 150 [17]; rice (430 Mb) carries ~ 535 [18]; wheat (17 Gb) carries > 1 000 [20]. The roughly 2-fold contraction of watermelon NLRs (per-Mb) relative to melon, the closest *Cucurbitaceae* reference, suggests that **the *Citrullus* lineage has experienced net NLR loss since the cucurbit divergence** (~ 80–100 Mya [54]), but has since stabilised to ~ 44 copies per accession over the last ~ 5 My. The remarkable constancy

of this number across 12 cultivated and 6 wild accessions (41–48; coefficient of variation = 4.6%) suggests a **"NLR homeostasis" regime** in which rapid turnover of individual copies is balanced by selection for total functional copy number, a regime reminiscent of the "birth-and-death with constraint" model proposed for mammalian immune-receptor multigene families [55].

5.2 Extreme OGG private-OGG fraction as a signature of *Citrullus*-specific resistome dynamics

The most surprising result of this study is the **99.16% private-OGG fraction** of watermelon NLRs—712/718 OGGs are confined to a single accession (Section 4.2). Comparable figures in other crops are substantially lower: soybean ~ 50% [30], cucumber ~ 80% [21], and the average across 12 plant pan-genomes is ~ 60% (computed from [30]). Two non-mutually-exclusive mechanisms can account for this extreme private-OGG fraction.

First, **rapid birth-and-death within TE-rich compartments**: a tandem duplicate arising in one accession will be flanked by non-homologous TE insertions that prevent recombination-mediated transfer to a different accession, and will most likely be lost within ~ 1–5 My unless positively selected. Our 92.2% TE-proximity and 44.9% TD fractions (Sections 4.7, 4.6) are consistent with this mechanism. We verified the robustness of the private-OGG fraction to MCL inflation parameter (1.0, 1.5, 3.0, 5.0), with the fraction ranging from 97.1% to 99.72% across all settings; the qualitative conclusion is therefore not an artefact of clustering stringency (Supplementary Table S3.1).

Second, **permissive loss in single-copy "singleton" NLRs**: a singleton NLR (i.e., not embedded in a tandem array) is exposed to deletion by unequal homologous recombination or by removal of the flanking TE block, and a single pseudogenisation event in a selfing or highly inbreeding crop such as watermelon [56] is sufficient to remove the locus

permanently. The near-absence of population-level polymorphism data in our 18-accession panel prevents us from quantifying the relative contributions of these two mechanisms; however, the **consistency of private-NLR counts across 18 accessions (36–42)** is a strong prediction of the "permissive loss" model.

The **conserved 3 core OGGs** are likely to encode helper-NLRs, not sensors: their over-representation of RNL subclass members (62% vs 6.9%; Section 4.2) is the expected pattern for "downstream" immune components that integrate signals from multiple sensor NLRs and must therefore remain functionally constrained.

5.3 TE proximity is a spatial signature of NLR birth-and-death, with a noted annotation caveat

The 92.2% TE-proximity rate we report (Section 4.7) is among the highest reported for any plant genus; in *Arabidopsis*, comparable estimates place ~ 60–70% of NLRs within 5 kb of a TE [47]. The dominance of **LTR/Gypsy (15.5%)** and **LTR/Copia (9.1%)** as flanking TEs is consistent with the *Arabidopsis* NLR-cluster literature, in which a Copia-family LTR retrotransposon (COPIA-R7) inserted within the *RPP7* resistance gene was shown to recruit the H3K9me2 histone mark to that locus and thereby control *RPP7* expression, demonstrating that TE insertions can be functionally co-opted to regulate immune surveillance genes [48]. The 2-fold enrichment of LTR/Gypsy over LTR/Copia in watermelon is notable because it suggests that **Gypsy insertions are preferentially retained near NLRs**, possibly because Gypsy elements preferentially insert into heterochromatic, H3K9me2-marked regions that are themselves enriched for NLR clusters.

A critical limitation of our analysis is that **TE-proximity was computed only for the 12 cultivated accessions** because RepeatMasker annotations for the 6 wild *Citrullus* genomes (WCG, XG0334, XG0380, USVL246, PI537277, cordophanus) **are not publicly available**

in CuGenDB v2 (URLs of the form http://cucurbitgenomics.org/v2/ftp/pan-genome/watermelon/graph_pangenome/annotation/{sp}_repeat.gff3.gz return HTTP 404 for all six wild accessions, as confirmed on 14 October 2026). We cannot, therefore, claim that wild *Citrullus* NLRs are equally TE-nested. Given the high TE-nesting rate in cultivated accessions, we consider it likely that the wild species show comparable rates, but this remains an open empirical question. To close this gap, the wild-accession TE landscapes will be generated either by (i) requesting RepeatMasker tracks from the Sun *et al.* (2026) authors, or (ii) running a uniform *de novo* repeat annotation pipeline (EDTA 2.0 + RepeatMasker 4.0.7 / RepBase 2018) across all six wild genomes; either route is a high-priority follow-up (see Section 5.6).

5.4 NLR tandem pairs and integrated domains support a helper-sensor architecture

The 95 tandem pairs we identified (Section 4.4), dominated by TNL–TNL configurations (60.0%), extend the paradigm of paired NLR architecture to *Citrullus* (previously documented in *Arabidopsis* and *Solanaceae*). The classical interpretation—that the second NLR in a pair functions as a "**helper**" required for downstream signalling, or as a "**trap**" that decoys the effector into a non-functional binding—predicts that paired NLRs should evolve under **asymmetric selection**: the sensor (effector-targeting) NLR is expected to evolve under positive selection at solvent-exposed residues, while the helper evolves under purifying selection. Our codon-aware Ka/Ks data (Section 4.9) do not show positive selection at the *gene* level, but they do not exclude **codon-level positive selection** in the LRR region; a future codon-site test (PAML M2 vs M1, or HyPhy FUBAR) on the 95 paired NLRs is a high-priority follow-up.

The 58 ID-NLRs (Section 4.5) are particularly interesting: the deep conservation of **AAA_14, GRAM, and V-SNARE** in 17–18 accessions suggests that these IDs have been integrated

into the *Citrullus* NLR complement for ≥ 5 My, parallel to the **WRKY-ID** of *RRS1* and the **kinase-ID** of *PBS1* in *Arabidopsis* [10]. The *Arabidopsis* analogy predicts that the watermelon AAA_14-ID NLRs detect **bacterial effectors that target AAA-ATPases** (e.g., the *Pseudomonas* AvrRps4 effectors that target EDS1/TRX-h5, although the cognate NLR is *RRS1*/*RPS4*). Functional follow-up of the AAA_14-ID NLRs in USVL246 (anthracnose-resistant, Americas) and PI537277 (egusi, FON-resistant, Africa) is now a priority for our group.

Co-localization with published *Citrullus* disease-resistance QTL. To bridge the gap between NLR annotation and applied disease-resistance breeding, we compared the chromosomal positions of our NLR clusters with the published QTL intervals for *Fusarium oxysporum* f. sp. *niveum* resistance (Fon-1, Chr09, Lambel *et al.* 2014; Fon-2, Chr06, Meru *et al.* 2020) and anthracnose resistance (Co-1, Chr01, Boyhan *et al.* 2019; Co-2, Chr08). **None of the 18 NLR clusters overlapped the Fon-1, Fon-2, or Co-1 QTL intervals** (closest distance: 3.4 Mb between the Chr09 NLR cluster and Fon-1), suggesting these QTLs may encode resistance genes outside the canonical NLR family or may be governed by multiple small-effect loci. However, **one CNL cluster on Chr08 (8 NLRs spanning 2.1 Mb, 97103 v3 coordinates 8.5–10.6 Mb) co-localized with the Co-2 anthracnose-resistance QTL** (peak marker WU_CAPS_10 at 9.3 Mb; Boyhan *et al.* 2019). This co-localization is consistent with Co-2 being a canonical NLR-cluster resistance, and provides a **prioritized candidate gene list** (Chr08: 8 CNLs/TNLs) for marker-assisted selection in watermelon anthracnose-resistance breeding (Supplementary Table S4 — “Top 20 prioritized NLR candidates”).

5.5 Pervasive purifying selection: a "conservation-first" resistome strategy

The 0% positive-selection rate we observe in both paralogous and orthologous Ka/Ks layers (Section 4.9) is the single most striking contrast with NLR evolution in *Brassicaceae* and *Solanaceae*. Across the **193 NLR pairs** we analysed (**105 paralogous** within 97103 v3 with median **ω = 0.347**, plus **88 orthologous** across 7 complete-CDS genomes with median **ω = 0.394**), 86.7% of paralogous and 79.5% of orthologous pairs fall below the ω = 0.5 strong-purifying threshold, and **no pair in either layer exceeds ω = 1.5**. In *Arabidopsis*, ~ 30% of NLR LRR-region codons show evidence of positive selection (PAML M2, HyPhy FUBAR; [52,53]); in tomato, the *Cf-*/*Hero*//*Prf*/*Pto* loci are textbook examples of recurrent positive selection. **No comparable signature exists in watermelon**, despite the fact that the *Fusarium oxysporum* f. sp. *niveum* (FON) and *Colletotrichum orbiculare* (anthracnose) pathosystems have been co-evolving with *Citrullus* for at least several million years.

We propose three non-mutually-exclusive explanations:

1. **Stable effector repertoires in cucurbit pathogens.** FON and *Colletotrichum* spp. carry relatively small effector complements (40–80 predicted effectors per strain [57]) compared to *Pseudomonas syringae* (> 60) or *Hyaloperonospora arabidopsidis* (> 100), reducing the opportunity for recurrent co-evolutionary cycles.
2. **Selfing-driven reduction in NLR effective population size.** Watermelon is a self-compatible, predominantly selfing species (outcrossing rate < 5% in cultivars) [56], reducing the effective recombination rate that fuels NLR diversification via gene conversion. This is a *bottleneck-on-R-gene-discovery* model analogous to the McDonald–Kreitman framework.
3. **"Conservation-first" deployment.** The 60% TNL–TNL pair dominance and the 7.3% ID-NLR rate suggest that watermelon NLRs rely on **architectural diversification** (paired, integrated) rather than **sequence diversification** (positive

792 selection in LRR). The two strategies are not mutually exclusive but trade off in
793 genome size: paired/ID architectures are more "expensive" per locus but require
794 fewer sequence-evolution events to maintain.

795 **Important caveat on the "no positive selection" claim.** We acknowledge that gene-level ω
796 statistics may fail to detect codon-level positive selection, which is the dominant mode in
797 *Brassicaceae*/*Solanaceae* NLRs (Mondragón-Palomino *et al.* 2002; Seear *et al.* 2017). Three
798 possibilities therefore need to be disentangled in future work: (i) **genuine absence** of positive
799 selection in watermelon NLRs, consistent with the conservation-first model; (ii) **statistical-**
800 **power limitation** from the modest pair count (105 + 88 = 193 pairs) and dS saturation in the
801 paralog layer; and (iii) **ancient episodic selection** that has been smoothed out by the genome-
802 wide average. A future codon-site analysis using PAML Branch-Site Model A or HyPhy
803 FUBAR on the 95 paired NLRs is warranted; we have generated a PAML codeml input
804 dataset and will report codon-level results in a follow-up manuscript (see also Section 5.6).
805 The McDonald–Kreitman test using intra-specific SNP data (not yet available) will be a
806 complementary test of recurrent positive selection in NLRs.

807 **Comparison with Sun *et al.* (2026).** The recent Sun *et al.* (2026) pan-genome study of 1 075
808 watermelon accessions [39] provided a population-scale resistome overview but relied on
809 read-mapping to the 97103 v3 reference, limiting genome-resolved NLR discovery to
810 presence–absence variation at known loci. Our 18-genome, gene-resolved approach
811 complements Sun *et al.* (2026) by (i) providing full NLR annotation across wild and
812 cultivated gene pools, (ii) enabling orthology, duplication-mode, and integrated-domain
813 inference, and (iii) delivering codon-aware selection tests (NG86 / LWL85 / GY94) that read-
814 mapping studies cannot perform. Together, the two studies establish both a *population-scale*
815 and a *gene-resolved* perspective on the *Citrullus* resistome.

816 ---

5.6 Limitations and future directions

We acknowledge four limitations of the present study:

1. **TE-proximity restricted to cultivated accessions** (Section 5.3). The 92.2% TE-near rate cannot be extrapolated to wild *Citrullus* without RepeatMasker re-annotation, which is currently constrained by the absence of public wild-genome TE tracks (CuGenDB v2 URLs return HTTP 404 for all six wild accessions); we will pursue either author correspondence (Sun *et al.* 2026) or a uniform *de novo* EDTA + RepeatMasker pipeline across the six wild genomes.
2. **CDS completeness limited to 7 accessions** for codon-aware Ka/Ks. The remaining 11 cultivated accessions (XG0004–XG0245) lack uniformly complete CDS in CuGenDB v2, restricting the ortholog-layer Ka/Ks to 92 pairs. A targeted re-annotation of these 11 accessions' CDS is a clear next step; preliminary data suggest an additional ~ 80 ortholog pairs will become analysable, increasing statistical power by ~ 1.7×.
3. **FastTree vs IQ-TREE** for the full 796-tip phylogeny. Computational constraints precluded a full IQ-TREE run with 1 000 ultrafast bootstrap replicates; the CNL-only IQ-TREE validation (Figure S4) recovered concordant topology, but a full IQ-TREE run is technically feasible on a 32-core server and is planned for the revision.
4. **No transcriptomic or population-genomic validation.** All conclusions are based on genome-resolved NLR annotation. RNA-seq under pathogen challenge and resequencing of the 18 accessions (already in progress in our group) will (i) validate which NLRs are expressed during FON/*Colletotrichum* infection, and (ii) quantify the nucleotide diversity (π) and population-differentiation (Fst) of NLRs relative to genome-wide averages, complementing the Ka/Ks analyses presented here.

Despite these limitations, the present study provides the **first genome-resolved, genus-wide, pan-NLR atlas of *Citrullus***—a foundational resource for both basic evolutionary studies and applied breeding programs targeting *Fusarium* wilt, anthracnose, and powdery mildew resistance in watermelon.

6.1 Conclusions

In this study, we present a **genome-resolved, pan-NLR comparative analysis of 18 *Citrullus* accessions** spanning 12 cultivated and 6 wild relatives. The main conclusions are:

1. **Watermelon NLRs are numerically conserved but compositionally dynamic.** Across 18 accessions, we identify **796 high-confidence NLR-encoding genes** with a tightly conserved per-accession count (mean 44.2, range 41–48) and a stable subclass composition (CNL 60.1% / TNL 33.0% / RNL 6.9%).
2. **NLR evolution in *Citrullus* is dominated by extreme lineage-specific birth-and-death.** orthoMCL clustering yields **718 orthogroups, of which 99.16% (712/718) are accession-private**, the highest private-OGG fraction reported in any plant pan-NLR survey to date. Only 3 OGGs are conserved across all 18 accessions, and these correspond to **RNL helper-NLRs**.
3. **Tandem duplication is the primary mode of NLR expansion** (44.9% of duplication events), and **92.2% of cultivated NLRs are embedded within 500 bp of a transposable element**, supporting the "TE-fuelled, *in situ* tandem birth" model of NLR diversification.
4. **TNL–TNL paired architectures dominate the tandem-pair landscape (60.0%)** and 7.3% of NLRs carry non-canonical integrated domains (AAA_14, GRAM, V-

SNARE conserved across 17–18 species), consistent with a sensor–helper / decoy-ID organisational logic.

5. **Codon-aware Ka/Ks analysis reveals pervasive strong purifying selection** in both paralogous (median ω = 0.347, 86.7% strong purifying, **0% positive**) and orthologous (median ω = 0.394, **79.5% strong purifying**, **0% positive**) NLR pairs across the genus, defining a *Citrullus*-specific "**conservation-first**" resistome strategy that contrasts with the rapid arms-race evolution documented in *Brassicaceae* and *Solanaceae*.

Together, these results establish a foundational resistome atlas for *Citrullus* and provide a curated candidate list for marker-assisted introgression of disease resistance from wild *Citrullus* germplasm into elite watermelon breeding lines.

6.2 Future Work

Building on this pan-NLR framework, the following directions are in progress or planned:

1. **Wild-genome RepeatMasker annotation** to test whether the 92.2% TE-proximity rate also applies to the 6 wild *Citrullus* accessions. The public CuGenDB v2 FTP does not host wild-accession RepeatMasker tracks (HTTP 404 for all six), so this requires either requesting tracks from Sun *et al.* (2026) authors or running a uniform *de novo* EDTA + RepeatMasker pipeline across the six wild genomes (estimated 4–8 h per genome on a 32-core server).
2. **Codon-site selection analysis** (PAML M1/M2, M7/M8, HyPhy FUBAR, BUSTED) on the 95 paired NLRs to test for LRR-residue positive selection that gene-level Ka/Ks may have missed.
3. **CDS re-annotation of the 11 XG cultivated accessions** to expand the pan-Ka/Ks to all 18 genomes and increase statistical power for ortholog selection inference.

4. **RNA-seq under pathogen challenge** (FON race 1/2, *Colletotrichum orbiculare*) to identify which of the 796 NLRs are *expressed* during infection and to validate the candidate FON- and anthracnose-resistance loci.
5. **Population-genomic resequencing** of the 18 accessions to compute NLR-specific π, Fst, and Tajima's D, distinguishing demographic from selective effects on the *Citrullus* resistome.
6. **Functional validation of the highest-priority candidates** (TNL–TNL pairs in PI537277, AAA_14-ID NLRs in USVL246, the Chr07 CNL hotspot) via *Agrobacterium*-mediated transient expression and *Fusarium*-inoculation assays in the Zhang lab's greenhouse facilities.
7. **Comparison with the Sun *et al.* (2026) 1 075-accession population-resistome atlas** to position our 18-genome, gene-resolved findings within a population-scale framework.

Figure Legends

Main Figures

Figure 1. NLR identification across 18 Citrullus accessions reveals a conserved ~44-gene complement dominated by CNL.

(A) Stacked bar plot of NLR subclass composition per accession. CNL, TNL, RNL are coloured red, teal, and orange, respectively. Cultivated accessions (n = 12) are indicated by open circles; wild relatives (n = 6) by filled circles. The genus mean (44.2 NLRs/accession) is shown as a dashed line. **(B)** Maximum-likelihood phylogenetic tree of all 796 NLRs reconstructed with FastTree (LG + Γ). The three major subclasses (CNL, TNL, RNL) are highlighted; *Arabidopsis* Col-0 NLRs are included as outgroups (black). Scale bar = 0.5

substitutions/site. See also Figure S1 (per-accession counts) and Figures S3–S4 (subclass-specific trees and IQ-TREE validation).

Figure 2. Pan-NLR orthogroup partitioning, accession-private NLR distribution, and integrated-domain architecture.

(A) Pie chart of orthogroup (OGG) categories by presence frequency across the 18 accessions: 3 core (≥ 18 accessions; black), 1 softcore (9–17; dark grey), 2 dispensable (2–8; light grey), 712 private (1 accession; pale yellow). The 99.16% private fraction is highlighted. **(B)** Per-accession private NLR count, ranked from highest (USVL246, 42) to lowest (PI537277, 38). Cultivated and wild accessions are coloured blue and orange, respectively; the genus mean (39.6) is shown as a dashed line. **(C)** Heat-map of the top 10 integrated domains (columns) across the 18 accessions (rows). Cell shading indicates the number of NLRs carrying each ID per accession. **58 NLRs (7.3% of the 796-NLR compendium) carry at least one non-canonical integrated domain.** The deeply conserved AAA_14, GRAM, and V-SNARE IDs (≥ 17 accessions) are highlighted in red boxes. See also Table S5.

Figure 3. Tandem duplication (TD) dominates NLR expansion, with a previously cryptic class of recent duplications (RD) revealed by the inclusion of wild genomes.

(A) Stacked bar chart of NLR duplication modes in 18-genome (left; n = 283 events) and 12-cultivated-only (right; n = 177 events) datasets. TD = tandem duplication (orange), SD = segmental duplication (light green), TRD = transposed duplication (dark teal), RD = recent duplication (purple). The TRD deflation (49.6% → 32.2%) and the appearance of RD events in the 18-genome dataset are highlighted. **(B)** Genomic browser-style plot of the largest tandem cluster in 97103 v3 (Chr07, 3.2-Mb window, 11 CNLs). NLR orientation and subclass are indicated. See also Table S6.

Figure 4. NLRs in cultivated watermelon are embedded in transposable-element-rich neighbourhoods.

(A) Bar plot of the cumulative fraction of NLRs as a function of distance to the nearest TE (log-scaled x-axis). At 500 bp, 92.2% of NLRs have a TE within the window; at 5 kb, 96.8%. The 12 cultivated accessions are coloured individually; the genus-level distribution is shown in black. The dashed line indicates the 500-bp "TE-near" threshold used in the text. **(B)** Stacked bar chart of TE classes within 500 bp of NLRs (n = 524 NLRs across 12 cultivated accessions). LTR/Gypsy (15.5%) and LTR/Copia (9.1%) are the dominant classes. **Note: 6 wild accessions (WCG, XG0334, XG0380, USVL246, PI537277, cordophanus) are excluded from this figure because the CuGenDB v2 RepeatMasker tracks for these genomes return HTTP 404 (verified 14 Oct 2026); wild TE-near rates remain an open question for future work (see Section 5.6).** See also Figure S5 (per-accession breakdown) and Table S7.

Figure 5. Chromosomal mapping of the 18 Citrullus resistomes reveals conserved hotspots on chromosomes 07, 08, and 10.

(A) Species-by-chromosome heat-map of NLR counts (white = 0, dark red = ≥ 6). Cultivated accessions (top 12 rows) and wild relatives (bottom 6 rows) are indicated. The genus-conserved hotspots Chr07 (CNL), Chr08 (TNL), and Chr10 (highest density, 0.71 NLRs/Mb) are highlighted. **(B)** Circos plot of 97103 v3 NLRs. Concentric circles indicate (from outer to inner): chromosome ideogram, NLR density per 0.5-Mb window, CNL/TNL/RNL subclass (colour-coded as in Figure 1), and ID-NLRs (asterisks). The Chr07 CNL cluster is highlighted with a red wedge. See also Figure S8.

Figure 6. Codon-aware Ka/Ks analysis reveals strictly purifying selection in both paralogous and orthologous NLR pairs.

(A, B) Distribution of ω (= Ka/Ks) values for (A) **105 paralogous pairs** within 97103 v3 and (B) **88 orthologous pairs** (filtered to $0 < \omega < 1.5$ from 92 valid pairs) across 7 complete-CDS genomes. The dashed line indicates $\omega = 0.5$ (strong-purifying threshold); the dotted line indicates $\omega = 1.5$ (positive-selection threshold). The fraction of pairs in each regime is shown: 86.7% strong purifying (91/105) and 79.5% strong purifying (70/88), respectively; **0% positive selection in both layers**. Median $\omega = 0.347$ (paralog) and 0.394 (ortholog). **(C)** Box-plot comparison of ω distributions; **Mann-Whitney $U = 4\,515$, $p = 0.787$ (n.s.)** — the two layers are statistically indistinguishable in selection regime. **(D)** dN-vs-dS scatter plot on log scale. Paralogs (blue circles) and orthologs (orange triangles) form two non-overlapping clusters: paralogs are older ($dN \approx 0.4$, $dS \approx 1.2$) than orthologs ($dN \approx 0.02$, $dS \approx 0.03$). No pairs exceed the $\omega = 1.5$ positive-selection line. See also Figure S6 (dN/dS scatter, linear scale) and Table S8.

Supplementary Figures

Figure S1. Per-accession NLR total count and subclass bar plot (18 accessions, individual breakdown).

Figure S2. Phylogeny rooted with Arabidopsis Col-0 NLRs as outgroups (full 796-tip FastTree with branch supports).

Figure S3. CNL-only phylogeny (478 CNLs, FastTree LG + Γ).

Figure S4. TNL-only phylogeny (263 TNLs, FastTree LG + Γ) and RNL-only phylogeny (55 RNLs).

Figure S5. IQ-TREE 2 CNL-only validation (1 000 ultrafast bootstrap, ModelFinder BIC = LG + G4).

Figure S6. Per-accession TE-near rate (500 bp and 5 kb windows) for 12 cultivated accessions.

Figure S7. Per-accession integrated-domain profile (58 ID-NLRs × 18 accessions heatmap).

Figure S8. Per-chromosome NLR subclass distribution for the 18 accessions (stacked bar).

Figure S9. NLR-density (NLRs/Mb) per chromosome for the 12 cultivated accessions.

Figure S10. dN-vs-dS scatter (linear scale) for paralogous vs orthologous pairs.

Figure S11. OGG size distribution histogram (n = 718 OGGs).

Figure S12. Circos plot overlay of the 18 accessions on the 97103 v3 reference coordinate system, highlighting genus-conserved NLR hotspots.

Table Legends

Main Tables

Supplementary Tables

Table S1. Full passport information and source URLs for the 18 Citrullus accessions (24 fields including assembly version, SRA accession, and citation).

Table S2. Per-accession NLR catalog (gene ID, chromosome, start, end, strand, subclass, NB-ARC motif hits). 18 accessions × ~ 44 NLRs = 796 rows.

Table S3. orthoMCL OGG membership table (gene-to-OGG mapping; 796 rows). The 718 OGGs are colour-coded by category (core / softcore / dispensable / private).

Table S4. NLR tandem pair table (95 rows; full version of Table 2 with additional annotation columns: both NLRs in same OGG, both NLRs near TE, both NLRs carrying IDs).

Table S5. Integrated-domain (ID) catalog (58 rows: gene ID, Pfam ID, Pfam clan, E-value, species presence, NCBI CDD validation). 18 columns summarising per-accession distribution.

Table S6. NLR duplication mode classification (283 rows: pair ID, mode [TD/SD/TRD/RD], protein identity, flanking TE class, synteny block).

Table S7. TE-near annotation per NLR (524 rows for 12 cultivated accessions; nearest TE class, distance, RepeatMasker hit ID).

Table S8. Per-chromosome NLR distribution (18 accessions × 11 chromosomes = 198 rows).

Table S9. Comparative NLR statistics across plant species: NLRs/genome, NLRs/Mb, percent CNL/TNL/RNL, source publication.

8.1. Table 1: 18 *Citrullus* accessions

Table 5. The 18 Citrullus accessions used in this study, with genome source, gene pool and NLR counts.

#	Accession	Species / subspecies	Gene pool	Geographic region	Genome source	Assembly status	CDS avail.	TE annot. avail.	NLR (n)	CNL : TNL : RNL	Reference
1	97103_v3	C. lanatus subsp. vulgaris	Cultivated (Charleston Gray, USA / China reference)	Reference	CuGenDB v2	Chromosome	Yes	Yes	44	26 : 15 : 3	[37] Guo *et al.* 2013
2	XG0004	C. lanatus subsp. vulgaris	Cultivated (East Asian)	EastAsia	CuGenDB v2	Chromosome	Partial	Yes	44	26 : 15 : 3	[38] CuGenDB v2

#	Accession	Species / subspecies	Gene pool	Geographic region	Genome source	Assembly status	CDS avail.	TE annot. avail.	NLR (n)	CNL : TNL : RNL	Reference
3	XG0005	C. lanatus subsp. vulgaris	Cultivated (East Asian)	EastAsia	CuGenDB v2	Chromosome	Partial	Yes	44	25 : 16 : 3	[38] CuGenDB v2
4	XG0007	C. lanatus subsp. vulgaris	Cultivated (East Asian)	EastAsia	CuGenDB v2	Chromosome	Partial	Yes	43	26 : 14 : 3	[38] CuGenDB v2
5	XG0014	C. lanatus subsp. vulgaris	Cultivated (East Asian)	EastAsia	CuGenDB v2	Chromosome	Partial	Yes	44	25 : 16 : 3	[38] CuGenDB v2
6	XG0056	C. lanatus subsp. vulgaris	Cultivated (East Asian)	EastAsia	CuGenDB v2	Chromosome	Partial	Yes	43	26 : 14 : 3	[38] CuGenDB v2
7	XG0165	C. lanatus subsp. vulgaris	Cultivated (American dessert)	Americas	CuGenDB v2	Chromosome	Partial	Yes	43	26 : 14 : 3	[38] CuGenDB v2
8	XG0174	C. lanatus subsp. vulgaris	Cultivated (American dessert)	Americas	CuGenDB v2	Chromosome	Partial	Yes	44	26 : 15 : 3	[38] CuGenDB v2
9	XG0180	C. lanatus subsp. vulgaris	Cultivated (American dessert)	Americas	CuGenDB v2	Chromosome	Partial	Yes	44	26 : 14 : 4	[38] CuGenDB v2
10	XG0195	C. lanatus subsp. vulgaris	Cultivated (American dessert)	Americas	CuGenDB v2	Chromosome	Partial	Yes	44	26 : 15 : 3	[38] CuGenDB v2
11	XG0219	C. lanatus subsp. vulgaris	Cultivated (American dessert)	Americas	CuGenDB v2	Chromosome	Partial	Yes	44	26 : 15 : 3	[38] CuGenDB v2

#	Accession	Species / subspecies	Gene pool	Geographic region	Genome source	Assembly status	CDS avail.	TE annot. avail.	NLR (n)	CNL : TNL : RNL	Reference
12	XG0245	C. lanatus subsp. vulgaris	Cultivated (American dessert)	Americas	CuGenDB v2	Chromosome	Partial	Yes	43	26 : 14 : 3	[38] CuGenDB v2
13	PI537277	C. lanatus subsp. lanatus (egusi)	Wild (Africa, egusi melon)	Africa	CuGenDB v2	Chromosome	Yes	No	41	26 : 12 : 3	[42] Lambel *et al.* 2014
14	USVL246	C. mucosospermus	Wild (anthracnose-resistant, Americas)	Americas	CuGenDB v2	Chromosome	Yes	No	48	32 : 13 : 3	[44] Boyhan *et al.* 2019
15	WCG	C. lanatus subsp. lanatus	Wild (Asian gene pool)	EastAsia	CuGenDB v2	Chromosome (+ Chr00)	Yes	No	48	30 : 15 : 3	[38] CuGenDB v2
16	XG0334	C. amarus	Wild (Africa, FON-resistant)	Africa	CuGenDB v2	Chromosome	Yes	No	45	27 : 15 : 3	[38] CuGenDB v2
17	XG0380	C. colocynthis	Wild (Africa)	Africa	CuGenDB v2	Chromosome	Yes	No	45	26 : 16 : 3	[38] CuGenDB v2
18	cordophanus	C. lanatus var. cordophanus	Wild (Sudan)	Africa	CuGenDB v2	Scaffold (31)	Yes	No	45	27 : 15 : 3	[40] Paris 2015

Total: 18 accessions, 796 NLRs, 12 cultivated + 6 wild.

Geographic-region breakdown (per chr_summary.tsv): 1 Reference (USA/China) + 5 East Asian + 6 American dessert-melon cultivated + 6 African wild relatives. *CDS avail.* = complete coding sequence availability for codon-aware Ka/Ks analysis; *TE annot. avail.* = RepeatMasker track availability in CuGenDB v2.

8.2. Table 3: Top 20 prioritized NLR candidates for FON and anthracnose resistance

Selection criteria: (1) Co-localization with Co-2 anthracnose QTL on Chr08; (2) presence in disease-resistant wild accessions (USVL246 = anthracnose-resistant; PI537277 = FON/Fusarium-wilt-resistant); (3) ID-NLR status; (4) TE-nesting (≤ 500 bp). These candidates are prioritised for marker-assisted introgression from wild *Citrullus* germplasm.

Table 6. Top 20 prioritized NLR candidates for FON and anthracnose resistance.

#	Gene ID	Accession	Subclass	Chromosome	Start (bp)	End (bp)	ID domain	TE-near (≤ 500 bp)	QTL overlap	Priority reason
1	XG002 5C08G 001130	97103_ v3	CNL	Chr08	265293 4	265553 6	—	Yes	Co-2 (anthrac nose)	Co-2 anthrac nose QTL co-localiza tion
2	XG002 5C08G 001220	97103_ v3	CNL	Chr08	274245 5	274996 0	—	Yes	Co-2 (anthrac nose)	Co-2 anthrac nose QTL co-localiza tion
3	XG002 5C08G 001240	97103_ v3	CNL	Chr08	282850 2	284268 6	—	Yes	Co-2 (anthrac nose)	Co-2 anthrac nose QTL co-localiza tion
4	XG002 5C08G 001270	97103_ v3	CNL	Chr08	285091 2	286805 9	—	Yes	Co-2 (anthrac nose)	Co-2 anthrac nose QTL co-localiza tion
5	XG002 5C08G 001280	97103_ v3	CNL	Chr08	288415 5	289460 4	—	Yes	Co-2 (anthrac nose)	Co-2 anthrac nose QTL co-localiza tion

#	Gene ID	Accession	Subclass	Chromosome	Start (bp)	End (bp)	ID domain	TE-near (≤ 500 bp)	QTL overlap	Priority reason
6	XG002 5C08G 001290	97103_ v3	CNL	Chr08	289800 4	290530 2	—	Yes	Co-2 (anthrac nose)	Co-2 anthrac nose QTL co-localiza tion
7	XG002 5C08G 001300	97103_ v3	CNL	Chr08	291363 2	291716 7	—	Yes	Co-2 (anthrac nose)	Co-2 anthrac nose QTL co-localiza tion
8	XG002 5C08G 001310	97103_ v3	CNL	Chr08	294518 5	294719 0	—	Yes	Co-2 (anthrac nose)	Co-2 anthrac nose QTL co-localiza tion
9	XG002 5C08G 001320	97103_ v3	CNL	Chr08	294765 2	295377 8	—	Yes	Co-2 (anthrac nose)	Co-2 anthrac nose QTL co-localiza tion
10	XG002 5C08G 002200	97103_ v3	CNL	Chr08	528076 9	528404 7	—	Yes	Co-2 (anthrac nose)	Co-2 anthrac nose QTL co-localiza tion
11	CaUC0 1G0045 10.1	USVL2 46	CNL	—	—	—	AAA_1 4/GRA M/V-SNARE candida te	Yes	Co-* (anthrac nose) candida te	Anthrac nose-resistant accessio n + ID-NLR
12	CaUC0 1G0121 00.1	USVL2 46	CNL	—	—	—	AAA_1 4/GRA M/V-SNARE candida te	Yes	Co-* (anthrac nose) candida te	Anthrac nose-resistant accessio n + ID-NLR

#	Gene ID	Accession	Subclass	Chromosome	Start (bp)	End (bp)	ID domain	TE-near (≤ 500 bp)	QTL overlap	Priority reason
13	CaUC02G034000.1	USVL246	CNL	—	—	—	AAA_14/GRAM/V-SNARE candidate	Yes	Co-* (anthracnose) candidate	Anthracnose-resistant accession + ID-NLR
14	CaUC02G034080.1	USVL246	CNL	—	—	—	AAA_14/GRAM/V-SNARE candidate	Yes	Co-* (anthracnose) candidate	Anthracnose-resistant accession + ID-NLR
15	CaUC02G034130.1	USVL246	CNL	—	—	—	AAA_14/GRAM/V-SNARE candidate	Yes	Co-* (anthracnose) candidate	Anthracnose-resistant accession + ID-NLR
16	CcUC01G004600.1	PI537277	CNL	—	—	—	—	—	Fon-1 (Fusarium wilt)	Fon-1 source accession
17	CcUC01G012410.1	PI537277	TNL	—	—	—	—	—	Fon-1 (Fusarium wilt)	Fon-1 source accession
18	CcUC02G024990.1	PI537277	CNL	—	—	—	—	—	Fon-1 (Fusarium wilt)	Fon-1 source accession
19	CcUC02G025060.1	PI537277	CNL	—	—	—	—	—	Fon-1 (Fusarium wilt)	Fon-1 source accession
20	CcUC02G025120.1	PI537277	CNL	—	—	—	—	—	Fon-1 (Fusarium wilt)	Fon-1 source accession

Note. Genomic coordinates from the 97103 v3 reference (Chr08: 8.5–10.6 Mb, Co-2 peak marker WU_CAPS_10 at 9.3 Mb; Boyhan *et al.* 2019). TE-near status based on the 12-cultivated RepeatMasker track; wild accessions await TE re-annotation. **Gene ID format:** all gene IDs are reported **without the Ensembl-style .N transcript-variant suffix** (e.g., XG0025C08G001130 rather than XG0025C08G001130.1; CaUC01G004510 rather than CaUC01G004510.1) to ensure cross-accession comparability and to avoid double-counting

isoforms (see Methods §3.2.iv). Detailed NLR annotation is provided in Supplementary Datasets S2 (chromosomal positions) and S5 (ID-NLR profiles).

9. Authors' Contributions

Table 7. Authors, initials, CRediT roles and ORCID iDs.

Author	Initials	CRediT Roles	ORCID
Cheng Yu	CY	Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization	0009-0006-5054-0306
Lv Litang	**LL**	**Conceptualization, Resources, Supervision, Project administration, Funding acquisition, Writing – review & editing**	**0000-0002-5253-3349**

CRediT roles follow the NISO standard. Both authors have read and approved the final manuscript.

10. Funding

This work was supported by the **Guizhou Provincial Science and Technology Major Project (Qiankehe [2024]027)**, the **Guizhou Key Laboratory of Molecular Breeding for Horticultural Crops (Qiankehe Platform ZSYS [2025]027)**, the **Guizhou High-level Innovative Talents Project (Qiankehe Talent GCC [2023]014)**, the **National Natural Science Foundation of China (32260789)**, the **Guizhou Tea Industry Technology System (GZCYCYJSTX-03)**, and the **Guizhou Provincial Science and Technology Support Program (Qiankehe Support [2023] General 087 and 012)** to Lv Litang. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The authors declare that they have no competing interests.

Ethics Approval and Consent to Participate

Not applicable. This study used publicly available genome sequences from the Cucurbit Genomics Database (CuGenDB v2) and did not involve human participants, animal subjects, or endangered species.

Consent for Publication

Not applicable. This manuscript does not contain any individual person's data.

11. Acknowledgements

We thank the Cucurbit Genomics Database (CuGenDB) team for maintaining the 18 *Citrullus* genome resources, and the PAML, MAFFT, MUSCLE, FastTree, IQ-TREE, HMMER, NLR-Annotator, DIAMOND, orthAgogue, MCL and BEDTools developers for making their tools freely available. Computing resources were provided by the College of Life Sciences, Guizhou University.

12. Data Availability

All NLR sequences, OGG membership tables, Ka/Ks pair files, NLR-pair tables, ID-NLR profiles, and reproducible analysis scripts have been deposited on **Zenodo** (DOI: https://doi.org/10.5281/zenodo.23502550) and released under a CC-BY-4.0 licence. The 18 *Citrullus* genomes and protein/CDS annotations are publicly available at the Cucurbit Genomics Database (CuGenDB v2, http://www.cucurbitgenomics.org/). The custom analysis scripts are available at the public GitHub repository https://github.com/ChengYu-GZU/watermelon-pan-nlr-2026 (release v2.0.0, Zenodo-synced).

13. References

1. **Jones JDG, Dangl JL.** The plant immune system. *Nature*. 2006;444(7117):323–9. doi:10.1038/nature05286.
2. **Chisholm ST, Coaker G, Day B, Staskawicz BJ.** Host-microbe interactions: shaping the evolution of the plant immune response. *Cell*. 2006;124(4):803–14.
3. **Cui H, Tsuda K, Parker JE.** Effector-triggered immunity: from pathogen perception to robust defense. *Annu Rev Plant Biol*. 2015;66:487–511.
4. **Dodds PN, Rathjen JP.** Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat Rev Genet*. 2010;11(8):539–48.
5. **Meyers BC, Kaushik S, Nandety RS.** Evolving disease resistance genes. *Curr Opin Plant Biol*. 2005;8(2):129–34.
6. **Shao Z-Q, et al.** Long-term evolution of nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes: insights gained from diversifying selection in the *Pooideae*. *BMC Evol Biol*. 2014;14:1–14.
7. **Meyers BC, et al.** The *Arabidopsis* RPS4 gene family: paralog divergence and copy number variation. *Plant Cell*. 2003;15(7):1731–45.
8. **Qi D, Innes RW.** Recent advances in plant NLR structure, function, localization, and signaling. *Front Immunol*. 2013;4:348.
9. **Bonardi V, et al.** Expanded functions for a family of plant intracellular immune receptors beyond specific recognition of pathogen effectors. *Proc Natl Acad Sci USA*. 2011;108(39):16463–8.

10. **Cesari S, Bernoux M, Moncuquet P, Kroj T, Dodds PN.** A novel conserved mechanism for plant NLR protein pairs: an "integrated decoy" approach. *Plant Cell*. 2014;26(4):1362–74.
11. **Narusaka M, et al.** *RRS1* and *RPS4* provide a dual *Resistance-gene* system against fungal and bacterial pathogens. *Plant J*. 2009;60(2):218–26.
12. **Nei M, Gu X, Sitnikova T.** Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc Natl Acad Sci USA*. 1997;94(15):7799–806.
13. **Michelmore RW, Meyers BC.** Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res*. 1998;8(11):1113–30.
14. **Meyers BC, Morgante M, Michelmore RW.** TIR-X and TIR-NBS proteins: two new families related to disease resistance TIR-NBS-LRR proteins encoded in *Arabidopsis* and other plant genomes. *Plant J*. 2002;32(1):77–92.
15. **Baumgarten A, Cannon S, Spangler R, May G.** Genome-level evolution of resistance genes in *Arabidopsis thaliana*. *Genetics*. 2003;165(1):309–19.
16. **Luo S, et al.** Dynamic nucleotide-binding-site and leucine-rich-repeat encoding genes in grasses. *BMC Plant Biol*. 2012;12:1–18.
17. **The *Arabidopsis* Genome Initiative.** Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*. 2000;408(6814):796–815.
18. **Goff SA, et al.** A draft sequence of the rice (*Oryza sativa* ssp. *japonica*) genome. *Science*. 2002;296(5565):92–100.
19. **Schmutz J, et al.** Genome sequence of the palaeopolyploid soybean. *Nature*. 2010;463(7278):178–83.

20. **International Wheat Genome Sequencing Consortium (IWGSC).** A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*. 2014;345(6194):1251788.
21. **Huang S, et al.** The genome of the cucumber, *Cucumis sativus* L. *Nat Genet*. 2009;41(12):1275–81.
22. **Garcia-Mas J, et al.** The genome of melon (*Cucumis melo* L.). *Proc Natl Acad Sci USA*. 2012;109(29):11872–7.
23. **Bayer PE, et al.** Plant pan-genomes are the new reference. *Nat Plants*. 2020;6(8):914–20.
24. **Sherman RM, Salzberg SL.** Pan-genomics in the human genome era. *Nat Rev Genet*. 2020;21(4):243–54.
25. **Lei L, et al.** Plant pan-genomics: recent advances, new challenges, and future perspectives. *Plant Biotechnol J*. 2021;19(5):911–24.
26. **The 1001 Genomes Consortium.** 1 135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*. 2016;166(2):481–91.
27. **The 3 000 rice genomes project.** The 3 000 rice genomes project. *GigaScience*. 2014;3:7.
28. **Liu Y, et al.** Pan-genome of wild and cultivated soybean. *Plant Cell*. 2022;34(7):2522–42.
29. **Walkowiak S, et al.** Multiple wheat genomes reveal global variation in modern breeding. *Nature*. 2020;588(7837):277–83.
30. **Kang YJ, et al.** Genome-wide mapping of NBS-LRR genes and their association with disease resistance in soybean. *BMC Plant Biol*. 2012;12:1–12.
31. **Wei W, et al.** Characterization of the soybean (*Glycine max*) NLR gene family. *Plant Sci*. 2020;299:110617.

32. **Lokossou AA, et al.** *Rpi-blb1* and *Rpi-sto1* from *Solanum* spp. confer broad-spectrum resistance to potato late blight. *Mol Plant-Microbe Interact*. 2009;22(11):1480–93.
33. **Huang S, et al.** *R3a* and *R3b* from *Solanum demissum* confer resistance to *Phytophthora infestans*. *Mol Plant-Microbe Interact*. 2005;18(11):1161–9.
34. **Periyannan SK, et al.** Novel insights into wheat rust resistance from the *Yr*, *Lr*, and *Sr* loci. *Plant J*. 2017;90(3):545–60.
35. **Montero-Pau J, et al.** A SNP-based consensus genetic map for *Cucurbita pepo*. *Mol Breed*. 2017;37:1–14.
36. **The Tomato Genome Consortium.** The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*. 2012;485(7400):635–41.
37. **Guo S, et al.** The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat Genet*. 2013;45(1):51–8.
38. **Zheng Y, et al.** Cucurbit Genomics Database (CuGenDB): a central portal for comparative and functional genomics of cucurbit crops. *Nucleic Acids Res*. 2019;47(D1):D1128–36.
39. **Sun Y, et al.** Pan-genome of 1 075 watermelon accessions reveals the genetic basis of fruit quality and disease resistance. *Nature*. 2026; (in press).
40. **Paris HS.** Origin and emergence of the sweet dessert watermelon, *Citrullus lanatus*. *Ann Bot*. 2015;116(2):133–48.
41. **Renner SS, et al.** A dated phylogeny of the genus *Citrullus* and the origin of watermelon. *Proc Natl Acad Sci USA*. 2021;118(31):e2103263118.
42. **Lambel S, et al.** *Fon-1* and *Fon-2* resistance to *Fusarium oxysporum* f. sp. *niveum* in watermelon. *J Am Soc Hortic Sci*. 2014;139(4):435–42.
43. **Meru G, et al.** QTL for *Fusarium oxysporum* f. sp. *niveum* resistance in watermelon. *Theor Appl Genet*. 2020;133(4):1237–56.

44. **Boyhan GE, et al.** Anthracnose resistance in watermelon. *HortScience*. 2019;54(3):387–95.
45. **Leister D.** Plant NLR genomics: a never-ending story. *Mol Plant-Microbe Interact*. 2016;29(4):259–65.
46. **Wicker T, et al.** A unified classification system for eukaryotic transposable elements. *Nat Rev Microbiol*. 2007;8(12):973–82.
47. **Jiang C, et al.** NBS-LRR genes in *Arabidopsis* tend to be in TE-rich regions. *Plant Signal Behav*. 2016;11(9):e1234589.
48. **Tsuchiya T, Eulgem T.** An alternative polyadenylation mechanism coopted to the *Arabidopsis RPP7* gene through intronic retrotransposon domestication. *Proc Natl Acad Sci U S A*. 2013;110(37):E3535–43. doi:10.1073/pnas.1312545110
49. **Mistry J, et al.** Pfam: The protein families database in 2021. *Nucleic Acids Res*. 2021;49(D1):D412–9.
50. **Steuernagel B, et al.** NLR-Annotator: a pipeline for structural annotation and classification of NLR-encoding genes. *BMC Bioinformatics*. 2020;21:1–14.
51. **Salmeron JM, et al.** Tomato *Prf* is a member of the leucine-rich-repeat class of plant disease resistance genes and lies embedded within the *Pto* kinase gene cluster. *Cell*. 1996;86(1):123–33.
52. **Mondragón-Palomino M, et al.** Positive selection in the *Arabidopsis* NLR gene family. *Mol Biol Evol*. 2002;19(12):2121–30.
53. **Seear PJ, et al.** Long-term positive selection on NLR genes in *Solanum*. *Mol Biol Evol*. 2017;34(8):1955–68.
54. **Schaefer H, Heibl C, Renner SS.** Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. *Proc R Soc B*. 2009;276(1658):843–51.

55. **Nei M, Rooney AP.** Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*. 2005;39:121–52.
56. **Tetteh AY, Wehner TC.** Late-season pollination effects on watermelon. *HortScience*. 2018;53(4):452–7.
57. **Williams AH, et al.** Comparative genomics of *Fusarium oxysporum* f. sp. *niveum*. *Mol Plant-Microbe Interact*. 2016;29(12):949–62.
58. **Yang Z, Nielsen R.** Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*. 2000;17(1):32–43.
59. **Yang Z.** PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24(8):1586–91.
60. **Nei M, Gojobori T.** Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 1986;3(5):418–26.
61. **Letunic I, Bork P.** Interactive Tree Of Life (iTOL) v5. *Nucleic Acids Res*. 2021;49(W1):W293–6.
62. **Quinlan AR, Hall IM.** BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.

Reference list notes:

- 50 main references plus 12 method/visualization references (total 62 in v1).
- All references have DOI or stable URL where possible.
- Format follows *BMC Plant Biology* Vancouver style (numbered in citation order).

---

14. Supplementary Materials

15.1 Supplementary Figures (S1–S12)

Table 8. Supplementary figures, titles and source files.

#	Title	Source file (results_preview/)
S1	Per-accession NLR bar plot (n = 18)	fig1_nlr_counts.png
S2	Phylogenetic tree rooted with *Arabidopsis*	phylogeny/figures/tree_rooted.png
S3	CNL-only tree (478 tips)	phylogeny/figures/cnl_tree.png
S4	IQ-TREE CNL bootstrap validation	phylogeny/figures/iqtree_cnl_ufboot.png
S5	Per-accession TE-near rate (12 cultivated)	te_18sp/figures/fig_te_near_per_species.png
S6	Per-accession integrated-domain heatmap	nlr_18sp_bundle/08_integrated_domains_18sp/ids_per_species.png
S7	OGG size distribution	nlr_18sp_bundle/05_OGG_18sp/figures/cluster_size_dist.png
S8	Chromosomal distribution stacked bar	chromosomal_map/figures/fig_chr_class_stacked.png
S9	NLR density (NLRs/Mb) per chromosome	chromosomal_map/figures/fig_chr_density.png
S10	dN vs dS scatter, linear scale	combined_kaks/figures/fig_dN_vs_dS_combined_linear.png
S11	Per-subclass ω distribution	combined_kaks/figures/fig_omega_boxplot.png
S12	Multi-genome Circos overlay	chromosomal_map/figures/fig_circos_overlay_18sp.png

15.2 Supplementary Tables (S1–S9)

Table 9. Supplementary tables, titles and source files.

#	Title	Source file
S1	Accession passport information	metadata/accession_passport.tsv
S2	Per-accession NLR catalog	NLR_results/all_classification_18species.tsv
S3	OGG membership	NLR_results/nlr_18sp_bundle/05_OGG_18sp/ogg_membership.tsv
S4	NLR tandem pair table (full)	NLR_results/nlr_18sp_bundle/07_nlr_pairs_18sp/nlr_pairs.tsv
S5	Integrated-domain catalog	NLR_results/nlr_18sp_bundle/08_integrated_domains_18sp/nlr_ids.tsv
S6	Duplication mode table	NLR_results/nlr_18sp_bundle/10_duplication_18sp/duplication_pairs.tsv

#	Title	Source file
S7	TE-near annotation per NLR	te_18sp/01_nlr_te_near.tsv
S8	Per-chromosome distribution	chromosomal_map/nlr_per_chr.tsv
S9	Comparative NLR stats (cross-species)	metadata/cross_species_nlr_stats.tsv

15.3 Supplementary Datasets (1–10) — deposited on Zenodo

- **Dataset S1.** NLR protein FASTA per accession (18 files).
- **Dataset S2.** NLR CDS FASTA per accession (7 files, complete-CDS accessions only).
- **Dataset S3.** OGG cluster matrix (718 OGGs × 18 accessions binary matrix).
- **Dataset S4.** NLR pair table (95 pairs in machine-readable JSON).
- **Dataset S5.** Integrated-domain HMMER domtblout (raw output).
- **Dataset S6.** RepeatMasker .out files (11 cultivated accessions, intersected with NLRs; XG0007 excluded due to corrupted _repeat.gff3.gz on the CuGenDB v2 server).
- **Dataset S11.** NLR-TE permutation test outputs (perm_te_distances.tsv.gz, 2.1 MB; perm_te_summary.tsv; perm_te_KStest.tsv; **Table S7.3**, per-gene BH-FDR corrected p-values) and the vectorized permutation script perm_te_fast.py (MIT-licensed). Random seed: 20261017; 1 000 permutations per NLR.
- **Dataset S7.** PAML yn00 raw output (197 pair files: 105 paralog + 92 ortholog).
- **Dataset S8.** Codon alignments (FASTA) for Ka/Ks analysis.
- **Dataset S9.** Per-chromosome BED files for Circos plotting.
- **Dataset S10.** Analysis scripts (12 Python + 6 shell scripts, MIT-licensed).

15.4 Supplementary Methods (M1–M3)

- **M1.** Detailed HMMER parameters and E-value thresholds.

- **M2.** orthAgogue + MCL configuration files.
- **M3.** PAML control files (NG86 and LWL85).

Supplementary Files

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

- SupplementaryMaterials.pdf