Pan-genomic insights into RLK family evolution and adaptation in Dioscorea alata

Zhi-Yan Wei
Nanjing Agricultural University
Sai-Xi Li
Nanjing University
Ming-Han Li
Jiangsu Province and Chinese Academy of Sciences
Jie Tang
Crop Research Institute of Jiangxi Academy of Agricultural Sciences
Yu-Qian Jiang
Nanjing University
Xin-Yu Lu
Jiangsu Province and Chinese Academy of Sciences
Yan-Mei Zhang
Jiangsu Province and Chinese Academy of Sciences
Jia-Yu Xue
xuejy@njau.edu.cn

Nanjing Agricultural University

Research Article

Keywords: Dioscorea alata, pan-RLKome, duplication, subfamiliy, evolution, adaptation

Posted Date: July 29th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7198655/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Plant Molecular Biology on October 20th, 2025. See the published version at https://doi.org/10.1007/s11103-025-01647-w.

Abstract

Dioscorea alata (greater yam) is a vital tuber crop underpinning global food security, while this crop suffers great reduction due to diseases like anthracnose, yam mosaic virus, and tuber rot. Receptor-like kinases (RLKs) play pivotal role in plant disease resistance, transducing plant immune signal. Yet the evolutionary dynamics of RLKs remain underexplored in *D. alata*. Here, we leveraged seven chromosome-level *D. alata* genomes to characterize the pan-RLKome of *D. alata*, identifying 4,119 RLK genes across 48 subfamilies. Our analysis revealed moderate variation in total RLK numbers but striking subfamily-specific expansions on several chromosome hotspots driven primarily by whole-genome/segmental duplication. Phylogenomic reconstruction uncovered pervasive extracellular domain swaps, fusions, and losses, contributing to diversified RLK architectures. Selection pressure analyses showed that purifying selection has maintained core RLK functions, while positive selection drove adaptive evolution in stress-associated subfamilies. Our study uncovers the pan-genomic basis of RLK evolution in *D. alata*, highlighting how duplication mechanisms and domain plasticity underpin functional innovation and environmental adaptation in this vital crop.

Key Message

Dioscorea alata's pan-RLKome reveals subfamily-specific expansions and domain plasticity driven by duplications, with purifying and positive selection shaping immune receptor evolution for disease adaptation.

Introduction

Greater yam (*Dioscorea alata*), also called water yam or winged yam, is cultivated worldwide, serving as an important food crop for millions of people especially in tropical and subtropical regions[1,2]. *D. alata* stands out in a number of cultivated yams (*Dioscorea* spp.) for its high production, ease of storage, and strong disease resistance[3,4]. The tubers of *D. alata* are rich in a variety of nutrients, including carbohydrates, proteins, fats, and vitamins[5]. The unique bioactive molecule relating to the antioxidant system, i.e. diosgenin, also provide *D. alata* particular properties as herb medicine[6]. As the world's fourth most important tuber crop (right behind cassava, potato, and sweet potato), *D. alata* has gained more and more attention for providing food diversity and security[1].

Signal perception and processing via cell-surface localized receptors is vital for all living organisms. In plants, receptor-like kinases (RLKs) function as receptors at the cell surface and represent one of the largest gene families[7,8]. Plant RLKs share a monophyletic origin with animal interleukin-1 receptor-associated kinase and Pelle kinases[9]. During plant evolution, the RLK family has experienced dramatic expansion and diversification, which is thought to be crucial for plant adaptation to changeable environments[10]. The domain architecture of RLKs is represented by the presence of a cytosolic kinase domain (KD), a transmembrane domain (TMD), and an extracellular domain (ECD)[7]. RLKs perceive both

self- and non-self-derived signals through diverse ECDs and transduces the signals downstream via their KDs[7]. Dozens of RLK subfamilies have been established based on the phylogenetic relationships of KD and the identity of ECD, with leucine-rich repeat (LRR)-RLK being the predominant subfamily[9,10,11]. It has been widely reported that RLKs can play roles in a diverse range of biological processes, including growth and development, reproduction, interaction with microbes, as well as biotic and abiotic responses[7,8]. For example, two well characterized LRR-RLKs *brassinosteroid insensitive 1* (*BRI1*) and *flagellin sensing 2* (*FLS2*) participate in brassinosteroid-mediated growth responses and bacterial flagellin perception, respectively[12,13,14,15].

In light of the recent publication of sequenced genome of *D. alata*[16], genome-wide identification and analysis of its RLK gene family has been conducted, providing useful insights into the evolutionary history and potential functional mechanism of RLK genes in *D. alata*[17]. However, with the great advancements in sequencing technology, more and more evidence show that a single “reference” genome is far from sufficient for fully depicting the genome content of one species[18]. Therefore, the concept of pan-genome has been proposed, meaning the whole non-redundant collection of genomic sequences of a species[18]. Pan-genome consists of the core genes, shared by all individuals, and the variable genes or dispensable genes, which are present only in some but not all individuals[19,20]. Dissecting the pan-genome is of great necessity for providing the information of genetically structural variants (SVs), including single-nucleotide polymorphisms (SNPs), insertions and deletions (InDels), copy-number variants (CNVs), and presence/absence variants (PAVs). These SVs have been proven to be the contributor to many important phenotypic variations[19,20]. However, the elaborate profile of *D. alata* RLK gene family in the context of pan-genome remains to be explored, impeding the comprehensive understanding on the evolutionary dynamics of RLK genes in *D. alata*.

In this study, with the aid of seven high-quality chromosome-level genomes of *D. alata*[21], we aimed to dissect the evolutionary dynamics of the RLK gene family at the pan-genome scale. By integrating genome-wide identification, phylogenetic reconstruction, and selection pressure analysis, we systematically explored the variation in RLK gene content, chromosomal distribution, duplication mechanisms, and domain architecture diversification across diverse accessions. Our work not only provides the first comprehensive characterization of the *D. alata* pan-RLKome but also unravels how specific RLK subfamilies have been shaped by both purifying and positive selection, reflecting their roles in balancing functional conservation and adaptive innovation. This pan-genome perspective bridges the gap between structural variation and functional evolution, offering a foundation for understanding RLK-mediated adaptability in *D. alata* and informing strategies for enhancing stress resilience in this vital crop species.

Methods

Genomes used in this study

Genomes of seven *D. alata* accessions were included in this study, which are ZSDa006 (guangzhougaotie)[21], ZSDa005 (suyu6hao), ZSDa015(ziyushanyao), ZSDa018 (S10 zisejiaobanshu), ZSDa019 (S22 jiaobanshu), ZSDa022 (S36 tangbeijiaobanshu), and ZSDa046 (nanyuanshenshu). The RLK sequences of the seven accessions are available at the China National GeneBank DataBase (CNGB) under accession number CNP0007639.

Identification of RLK genes

The primary transcripts of the annotated proteomes across seven *D. alata* accessions were filtered for downstream analysis. First, the hmmsearch programme in HMMER (v.3.3)[26] was used to search the KD (Pkinase, Pfam: PF00069.28) (e-value: 1×10^{-1}). Second, we scanned all domains presented in the obtained kinase candidates through the hmmscan programme in HMMER (v.3.3)[26] (e-value: 1×10^{-1}). Third, TMD and potential signal peptide were identified by adopting TMbed, a protein language model-based method in predicting the transmembrane helix or sheet domain[22]. Proteins with N-terminal ECD, intermediate TMD and C-terminal KD were identified as RLKs.

Classification of RLK genes

Dozens of RLK subfamilies have been established based on the phylogenetic relationships of KD and the identity of ECD[11]. The hidden Markov models (HMMs) built in a previous landmark study[11] were used to assign RLK members into different subfamilies and subgroups through hmmsearch (e-value: 1×10^{-10}).

Duplication events and collinearity analysis

On the basis of copy number and genomic distribution, genes within a single genome can be classified as one of the five types: singleton, dispersed duplicate, proximal duplicate, tandem duplicate and segmental/WGD duplicate[27]. We initially conducted all-against-all blast analysis for protein sequences of each seven *D. alata* accession via DIAMOND (v.2.0.14)[28] (e-value: 1×10^{-5}). Next, the obtained blast result files and genome annotation files were used to determine duplication type and collinearity relationship of all annotated proteins through MCScanX[23]. The “advanced circus” module of TBtools-II[27] was adopted to visualize the synteny blocks of RLK genes.

Sequence alignment and phylogenetic analysis

Depending on the hmmscan results (e-value: 1 × 10 − 1), we extracted the protein sequences of KDs from RLKs. Then, we adopted MAFFT (v.7)[29] to align the KD sequences. FastTree (v.2.1.11)[30] was used to conduct the phylogenetic analyses with default parameters. The R package ggtree (v.3.16)[31] was used to visualize the phylogenetic trees.

Orthogroup determination and selection pressure analysis

Clades were defined as OGs if all descendant leaves had cumulative branch lengths ≤ 0.05 and represented ≥ four accessions. Non-overlapping clusters were prioritized by size, with remaining genes classified as non_OG. We performed the selection pressure analysis on each determined OG through ParaAT (v.2)[32] and KaKs_Calculator 2.0 (ref.[33]).

Results

Construction of D. alata pan-RLKome based on genomes from seven accessions

Seven *D. alata* accessions with chromosome-level genomes were collected in this study, one of which was assembled to near telomere-to-telomere (T2T) standard[21]. By utilizing the language-model-based RLK detection method[17,22], we identified and classified the RLK gene family in each *D. alata* accession. A total of 4,119 RLK genes were retrieved, with 3,552 RLKs possessing both ECDs and TMDs, and 568 RLKs having only TMDs but lacking definite ECDs (Fig. 1A; Supplementary Table 1; Supplementary Data 1). Considering that the annotated ECDs were not deemed as the indispensable criterion for identifying RLKs in recent researches, we retained those RLKs without ECDs to maximize the prediction of transmembrane receptors. The average RLK gene number of seven *D. alata* accessions was 588. ZSDa018 and ZSDa019 had the most and fewest RLK genes among them, with number of 603 and 577, respectively (Fig. 1A). This indicated that the variation degree of *D. alata* RLK gene number was moderate.

We grouped RLK genes into 48 subfamilies by interrogating plant kinase Hidden Markov Models (HMMs) established by phytogenic relationships in a previous landmark study[11]. The largest subfamily was LRR-RLK as expected, with its number ranging from 230 in ZSDa019 to 243 in ZSDa015, followed by DLSV (DUF26, SD-1, LRR-VIII, and VWA), L-LEC, SD-2b, and LRK10L-2 (Fig. 1B). On the basis of HMMs, LRR-RLK were further classified into 23 subgroups, among which LRR-XI-1, LRR-III, and LRR-XII-1 being the first three abundant subgroups (Fig. 1C). Subfamilies not mentioned above, with or without LRR, tended to have comparatively smaller but more constant gene number (Fig. 1C-D).

To determine the variation degree of *D. alata* RLK subfamily gene number, we calculated the standard deviation of each subfamily. Our results showed that LRK10L-2, LRR-XI-1, DLSV, LRR-III, and SD-2b exhibited the highest dispersion degree (Supplementary Fig. 1). Interestingly, they represented the biggest subfamilies at the same time (Fig. 1B-D). By comparison, the small subfamilies were often more constant and conserved since their values of standard deviation were lower and some even reached 0 (Supplementary Fig. 1). Since the large RLK subfamilies are often reported to be versatile, their flexibility might reflect the capability of plant adaptation throughout evolution. Through scrutinizing of our data, we observed a remarkable increase of the LRK10L-2 gene number in ZSDa018, with at least 20 more than others, which could be the cause of the larger RLK gene number in ZSDa018 as well as the higher dispersion degree in LRK10L-2 (Fig. 1D; Supplementary Table 1).

D. alata pan-RLKome localization dynamics underscores the chromosomal hotspots

To explore the chromosomal distribution pattern of *D. alata* RLK genes, we retrieved their physical locations from each *D. alata* accession. Our results showed that RLK genes were distributed in all 20 *D. alata* chromosomes, among which chromosome (Chr) 15 (Chr15) contained the most RLK genes (range from 82 to 93), whereas only two RLK genes were distributed on Chr6 (Fig. 2A). Interestingly, RLK gene number among seven *D. alata* accessions maintained constant only on two chromosomes (Chr6 and Chr11), but showed more or less variable on other chromosomes (Fig. 2A; Supplementary Fig. 2). We found the largest value of standard deviation on Chr16, where ZSDa018 owned a substantially increase of RLK (Fig. 2A; Supplementary Fig. 2). Collectively, these results suggested that both the inter- and intra-chromosomal distributions of RLK genes were uneven and swiftly fine-tuned throughout evolution.

Considering the complex composition of RLK gene family, we inquired the arrangement of each RLK subfamily on *D. alata* chromosomes and analyzed their inter-accession differences. The results showed that Chr5 had the most diversified RLK subfamilies (23 subfamilies), while Chr6 had only two RLK subfamilies (LRR-III and RLCK-V) (Fig. 2B). Interestingly, although the most widely distributed subfamily, LRR-III, could be detected on 18 chromosomes, a total of 10 subfamilies were localized on merely one chromosome (Fig. 2B). Furthermore, the value of standard deviation of LRK10L-2 on Chr16, LRR-XI-1 on Chr2, and DLSV on Chr15 ranked at the top three, which exactly coincided with the most variable subfamilies as well as chromosomes (Fig. 1C-D; Fig. 2B; Supplementary Figs. 1–2). Taken together, our resulted indicated that the variation of pan-RLKome in *D. alata* was not unlimited and could be attributed to differences caused by members from specific RLK subfamilies and on specific chromosomes.

Diverse duplication mechanisms shaped the expansion history of D. alata pan-RLKome

In view of the large size of RLK gene family, we delved into its expansion history in *D. alata* under the background of pan-genome. The duplication types of RLK genes from seven *D. alata* accessions were analyzed by adopting MCScanX package[23] (Supplementary Table 2). Approximately a quarter of RLK genes were generated by tandem duplication, and another quarter were resulted from whole genome duplication (WGD) and/or segmental duplication (SD) (Fig. 3A). In addition, less than 20% RLK genes were resulted from proximal duplication (Fig. 3A). It was noteworthy that no singleton RLK gene was detected in all seven *D. alata* accessions (Fig. 3A), indicating that RLK repertoires were largely shaped by the recurrent and ubiquitous duplication events.

We further investigated the duplication types of each RLK subfamily and found that the majority of RLK subfamilies have experienced dispersed duplication (44/48) and WGD/SD (30/48) (Fig. 3B). By comparison, tandem duplication (15/48) and proximal duplication (14/48) contributed to the expansion of only one third of RLK subfamilies, respectively (Fig. 3B). Interestingly, RLK genes duplicated via WGD/SD were more abundant in ZSDa018 and ZSDa046 (Fig. 3A). Compared to other accessions, ZSDa018 and ZSDa046 contained remarkably elevated amount of WGD/SD-derived LRK10L-2 and LRR-XI-1 genes, respectively, and both of them also possessed larger number of WGD/SD-derived DLSV genes (Fig. 3B; Supplementary Fig. 3).

To dissect the expansion dynamics of RLK subfamilies in *D. alata*, we conducted the synteny analysis and delineated the specific gene loci that were produced by WGD/SD. Our analysis revealed extensive blocks of collinearity within the genomes of seven accessions, pinpointing numerous duplicated regions derived from ancient polyploidization and/or large-scale segmental duplication events (Supplementary Fig. 4A-G). Specifically, in ZSDa018, clusters of WGD/SD-derived LRK10L-2 genes were predominantly localized within collinear blocks on Chr16 (Supplementary Fig. 4D), indicating their origin from shared ancestral duplication events. Similarly, ZSDa046 exhibited dense clusters of LRR-XI-1 genes within syntenic regions on Chr2 (Supplementary Fig. 4G). Furthermore, the enrichment of DLSV genes via WGD/SD in both accessions was corroborated by their frequent occurrence within these identified collinear blocks across Chr5 (Supplementary Fig. 4A-G). These intra-specific synteny patterns provide strong genomic evidence supporting the significant role of WGD/SD events in driving the lineage-specific expansion of key RLK subfamilies.

Evolutionary trajectory and domain architecture diversification of D. alata pan-RLKome

The intricate evolutionary trajectory of RLK gene family has long been mysterious due to its large size and numerous subfamily composition. Several dedicated researches have been conducted for model plants[9,10,11,17], but the pan-genome-based analysis still remains scarce. To provide a panoramic evolutionary history of *D. alata* RLK gene family, we reconstructed the phylogeny of the pan-RLKome in seven *D. alata* accessions (Supplementary Fig. 5; Supplementary Data 2). Interestingly, we found that one same clade in the phylogenetic tree could often accommodate RLKs derived from different subfamilies and/or located on different chromosomes, indicating the frequent ECD changes and chromosome-translocations of RLKs (Fig. 4A). For example, as the largest RLK subfamily, LRR-RLK did not form a monophyletic group but rather belonged to several different clades (Fig. 4A). Similarly, subfamilies of receptor-like cytosolic kinases (RLCKs) were also distributed on a couple of clades (Fig. 4A).

A wide variety of types of domains are integrated into the RLK gene family to serve as ECDs. Therefore, we probed into the evolutionary roadmap of RLK domain architectures by determining the ECD types. Totally, eight classes of ECD type were identified, including cysteine-rich repeat (CRR), lectin, LRR, lysin motif (LysM), malectin, regulator of chromosome condensation repeat (RCC), similar protein to arbuscular receptor-like kinase (SPARK), and wall-associated kinase (WAK) (Fig. 4B; Supplementary Figs. 6–7; Supplementary Table 3). CRR, LysM and RCC domains were specifically adopted as ECDs in DLSV, LysM-RLK and CR4L subfamilies, respectively (Fig. 4B; Supplementary Figs. 6–7). Among the subfamilies with lectin as ECD, C-LEC, L-LEC, and SD-2b/DLSV held the C-type, L-type, and B-type lectin domains, respectively (Fig. 4B; Supplementary Figs. 6–7). LRR, malectin, SPARK, WAK domains were integrated into more than one RLK subfamilies, underscoring the versatility of these ECDs (Fig. 4B; Supplementary Figs. 6–7). Despite the various ECD types, no ECD was found at all in 13 RLK subfamilies, which implied their intrinsic role in intercellular signal transduction but not perceiving extracellular stimuli (Fig. 4B; Supplementary Fig. 6–7).

The phylogeny of RLK subfamilies mapped with ECD types showed the swift and frequent ECD domain change events (Fig. 4B). The inositol-requiring enzyme 1 (IRE1) clade that contained TMD but no ECD was the sister group to all other RLK subfamilies (Fig. 4B). Homologs of IRE1 have been found in human, suggesting the ancient origin of this transmembrane kinase receptors[24]. Notably, LRR domains were predominant in many clades (Fig. 4B), which raised the possibility that other domains could be recruited into the RLK through ECD swap mediated by LRR. Intriguingly, 14 subfamilies that comprised mostly of RLKs with ECD simultaneously contained RLKs without ECD (Supplementary Figs. 6–7). The pervasive distribution of subfamilies with no specific ECDs were across the phylogenetic tree of RLKs highlighted the universality of the ECD loss (Fig. 4B). In a nutshell, the prevailing ECD fusion, swap, and loss events shaped the domain architectures of the RLK gene family throughout the evolution.

Purifying and positive selection jointly shaped functional specialization of RLK subfamilies in D. alata

To analyse the evolutionary dynamics of the *D. alata* pan-RLKome, we determined orthologous groups (OGs) of RLK genes using a post-order traversal strategy through systematically scanning internal nodes of the phylogenetic tree (detailed in Methods) (Supplementary Fig. 5; Supplementary Data 2). We identified 493 OGs containing totally 3,900 RLK genes, accounting for more than 94% of the *D. alata* pan-RLKome (Fig. 5A; Supplementary Fig. 8; Supplementary Table 3). Most of the OGs (432/493, 87.6%) contained genes from all seven accessions (Fig. 5A; Supplementary Table 3). Conserved OGs that possessed no more than two genes per accession accounts for the vast majority (475/493, 96.3%) (Fig. 5A; Supplementary Table 3). This indicated that our approach allowed identification of evolutionarily conserved gene groups while maintaining accession-specific resolution.

To elucidate the evolutionary forces acting on the *D. alata* pan-RLKome, we calculated the ratio of nonsynonymous substitution rates (Ka) to synonymous substitution rates (Ks) (Ka/Ks) for orthologous RLK genes across seven accessions. Among the 493 OGs, 240 highly conserved OGs (48.6%) exhibited Ka = 0 and Ks = 0, indicating strong purifying selection and no detectable changes of the nucleotide sequences (Fig. 5B; Supplementary Table 4). For the remaining OGs, a pronounced proportion of them showed Ka/Ks values around 1, which meant that neutral selection and genetic drift served as the major contributor (Fig. 5B; Supplementary Table 4). Notably, many subfamilies harboured OGs under positive selection alongside those under purifying selection, such as SD-2b, L-Lec and DLSV, suggesting their versatility (Fig. 5C). By comparison, RLK subfamilies involved in conserved biological processes (e.g., LRR-V, WAK) showed remarkably lower Ka/Ks values (Fig. 5C). Further subgroup analysis revealed that RLKs derived from proximal duplication events possessed more OGs with Ka/Ks < 1 values compared to those from dispersed, tandem or WGD/SD duplications (Fig. 5D). Collectively, these results underscore the heterogeneous evolutionary pressures across RLK subfamilies, shaped by their functional specialization and duplication mechanisms.

Discussion

This study represents the first comprehensive dissection of the RLK gene family at a pan-genome scale in *D. alata*, leveraging seven high-quality chromosome-level assemblies. Our analysis revealed a pan-RLKome of 4,119 genes, extending beyond previous single-reference genome analyses which identified ~ 500 RLKs[17]. In addition, pan-genomic analysis reveals richer structural variations (e.g., copy number variations and presence-absence variations) that likely drive RLK functional diversification.

The moderate variation in total RLK numbers across accessions (577–603 genes) contrasts with striking subfamily-specific expansions, notably in LRK10L-2, LRR-XI-1, DLSV, and SD-2b. Synteny and duplication analyses demonstrate that WGD/SD serve as the primary drivers of these expansions, particularly on chromosomal hotspots, such as LRK10L-2 on Chr16 in ZSDa018 and LRR-XI-1 on Chr2 in ZSDa046 (Supplementary Fig. 4). This aligns with findings in other plants, where WGD/SD events facilitate rapid neofunctionalization and adaptation gene families[25]. Tandem and proximal duplications contributed to a smaller fraction of RLKs (15–25%), underscoring the dominant role of large-scale genomic reconfigurations in shaping RLK diversity (Fig. 3).

Phylogenomic reconstruction revealed pervasive ECD swaps, fusions, and losses across RLK clades. Subfamilies historically defined by conserved ECDs, such as LRR-RLKs, were polyphyletic, with domains like lectin, malectin, and LysM frequently exchanged or lost (Fig. 4A). This architectural fluidity suggests evolutionary flexibility in ligand perception strategies, potentially enabling novel environmental sensing capabilities. Notably, 13 subfamilies lacked identifiable ECDs, implying roles in intracellular signaling or co-receptor functions.

OG analysis identified 493 conserved RLK groups, 87.6% of which were shared across all accessions, highlighting a core RLKome stabilized by purifying selection. Strikingly, 48.6% of OGs showed no detectable sequence divergence (Ka = Ks = 0), indicating intense functional conservation (Fig. 5). Conversely, RLK subfamilies associated with stress responses (e.g., SD-2b, L-LEC, and DLSV) exhibited signatures of positive selection (Ka/Ks > 1), suggesting adaptive evolution to biotic/abiotic challenges. This dichotomy aligns with dual roles of RLKs: conserved kinases maintain developmental integrity, while stress-linked clades diversify under environmental pressures[10]. Subfamilies derived from proximal duplications showed stronger purifying selection than those from WGD/SD, implying tighter functional constraints in localized genomic contexts.

The pan-RLKome landscape provides insights into *D. alata* adaptive evolution, particularly its stress resilience. The expansion of stress-related RLK subfamilies (e.g., LRK10L-2, DLSV) and the plasticity of ECDs likely contribute to its strong resistance traits, a key agronomic advantage. These findings offer targets for molecular breeding. For example, leveraging natural variation in RLK copy numbers or domain architectures to enhance disease resistance or abiotic stress tolerance. Future studies could functionalize specific RLKs via gene editing to validate their roles in stress responses, or expand pan-genomic analyses to include wild Dioscorea species to uncover ancestral RLK functions.

In summary, our pan-genomic analysis reveals how duplication-driven expansion and domain plasticity have shaped the *D. alata* RLK repertoire, balancing functional conservation with adaptive innovation. This framework not only advances our understanding of plant receptor evolution but also provides a foundation for leveraging RLK diversity in crop improvement strategies.

Declarations

Author Contributions Statement

J.Y.X. and Y.M.Z. conceived and designed the research. Z.Y.W., S.X.L. and M.H.L. obtained and analyzed the data; S.X.L., M.H.L. and Z.Y.W. drafted the manuscript; J.T., Y.Q.J. and X.Y.L. participated in data analyses; J.Y.X. and Y.M.Z. revised the manuscript. All the authors read and approved the final manuscript.

Competing Interests Statement

The authors declare no conflict of interests.

Author Contribution

J.Y.X. and Y.M.Z. conceived and designed the research. Z.Y.W., S.X.L. and M.H.L. obtained and analyzed the data; S.X.L., M.H.L. and Z.Y.W. drafted the manuscript; J.T., Y.Q.J. and X.Y.L. participated in data analyses; J.Y.X. and Y.M.Z. revised the manuscript. All the authors read and approved the final manuscript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32172089), Scientific Fund of Nanjing Botanical Garden Mem. Sun Yat-Sen (JSPKLB202503).

References

1. Lebot V, Lawac F, Legendre L (2023) The greater yam (Dioscorea alata L.): A review of its phytochemical content and potential for processed products and biofortification. J Food Compos Anal 115
2. Asiedu R, Sartie A (2010) Crops that feed the World 1. Yams Food Secur 2:305–315
3. Salehi B et al (2019) Dioscorea Plants: A Genus Rich in Vital Nutra-pharmaceuticals-A Review. Iran J Pharm Res 18:68–89

4. Adomeniene A, Venskutonis PR (2022) Dioscorea spp.: Comprehensive Review of Antioxidant Properties and Their Relation to Phytochemicals and Health Benefits. Molecules 27
5. Zang Z, Gong X, Cao L, Ni H, Chang H (2024) Resistant starch from yam: Preparation, nutrition, properties and applications in the food sector. Int J Biol Macromol 273:133087
6. Jesus M, Martins APJ, Gallardo E, Silvestre S (2016) Diosgenin: Recent Highlights on Pharmacology and Analytical Methodology. J Anal Methods Chem 2016:4156293
7. Dievart A, Gottin C, Périn C, Ranwez V, Chantret N (2020) Origin and Diversity of Plant Receptor-Like Kinases. Annu Rev Plant Biol 71:131–156
8. Gish LA, Clark SE (2011) The RLK/Pelle family of kinases. Plant J 66:117–127
9. Shiu SH, Bleecker AB (2001) Receptor-like kinases form a monophyletic gene family related to animal receptor kinases. Proc Natl Acad Sci USA 98:10763–10768
10. Lehti-Shiu MD, Zou C, Hanada K, Shiu SH (2009) Evolutionary History and Stress Regulation of Plant Receptor-Like Kinase/Pelle Genes. Plant Physiol 150:12–26
11. Lehti-Shiu MD, Shiu SH (2012) Diversity, classification and function of the plant protein kinase superfamily. Philos Trans R Soc B 367:2619–2639
12. Snoeck S, Johanndrees O, Nürnberger T, Zipfel C (2025) Plant pattern recognition receptors: from evolutionary insight to engineering. Nat Rev Genet 26:268–278
13. Chen XX et al (2021) Protein kinases in plant responses to drought, salt, and cold stress. J Integr Plant Biol 63:53–78
14. Manhaes AMA, Ortiz-Morea FA, He P, Shan L (2022) Plant plasma membrane-resident receptors: Surveillance for infections and coordination for growth and development. J Integr Plant Biol 64:1132–1132
15. Cui YW, Lu XT, Gou XP (2022) Receptor-like protein kinases in plant reproduction: Current understanding and future perspectives. Plant Commun 3
16. Bredeson JV et al (2022) Chromosome evolution and the genetic basis of agronomically important traits in greater yam. Nat Commun 13
17. Jiang YQ, Lu XY, Qin YL, Zhang YM, Shao ZQ (2024) Genome-Wide Identification and Evolutionary Analysis of Receptor-like Kinase Family Genes Provides Insights into Anthracnose Resistance. Plants 13
18. Golicz AA, Batley J, Edwards D (2016) Towards plant pangenomics. Plant Biotechnol J 14:1099–1105
19. Shi JP, Tian ZX, Lai JS, Huang XH (2023) Plant pan-genomics and its applications. Mol Plant 16:168–186
20. Jayakodi M, Shim H, Mascher M (2024) What Are We Learning from Plant Pangenomes? Annu Rev Plant Biol
21. Zhang YM et al (2025) A telomere-to-telomere genome assembly for greater yam (Dioscorea alata). Plant Commun :101326

22. Bernhofer M, Rost B (2022) TMbed: transmembrane proteins predicted through language model embeddings. BMC Bioinformatics 23
23. Wang YP et al (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40
24. Chen Y, Brandizzi F (2013) IRE1: ER stress sensor and cell fate executor. Trends Cell Biol 23:547–555
25. Ren R et al (2018) Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms. Mol Plant 11:414–428
26. Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195
27. Chen C et al (2023) TBtools-II: A one for all, all for one bioinformatics platform for biological big-data mining. Mol Plant 16:1733–1742
28. Buchfink B, Reuter K, Drost HG (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368
29. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30:772–780
30. Price MN, Dehal PS, Arkin AP (2010) FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 5
31. Xu S et al (2022) Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. iMeta 1:e56
32. Zhang Z et al (2012) ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. Biochem Biophys Res Commun 419:779–781
33. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J (2010) KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. Genomics Proteom Bioinf 8:77–80

Figures

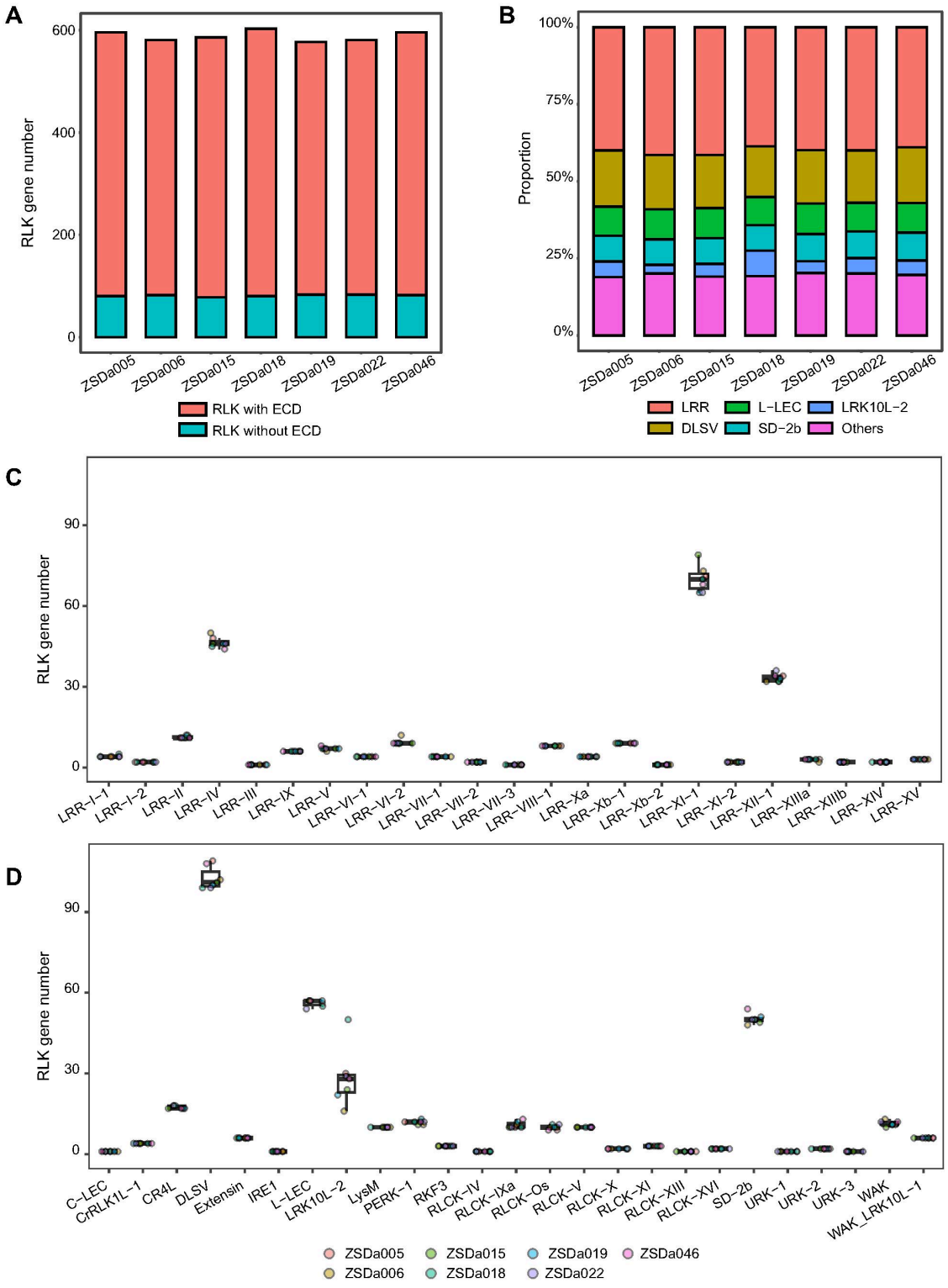

Figure 1

D. alata **pan-RLKome constructed on seven accession genomes. A** Distribution of the RLK gene numbers across seven *D. alata*accession genomes. The RLK gene numbers are presented as bar plots. RLK with ECD is indicated in red, while RLK without ECD is indicated in cyan. **B**Proportions of the RLK subfamilies in *D. alata* pan-RLKome. The proportions are presented as bar plots. LRR, DLSV, L-LEC, SD-2b, LRK10L-2, and other subfamilies are indicated in different colours. **C** Distribution of the RLK gene numbers across

subgroups of LRR-RLKs. The RLK gene numbers are presented as boxplots, with points representing each accession in different colours (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, each accession). **D** Distribution of the RLK gene number across non-LRR-RLK subfamilies. The RLK gene numbers are presented as boxplots, with points representing each accession in different colours (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, each accession).

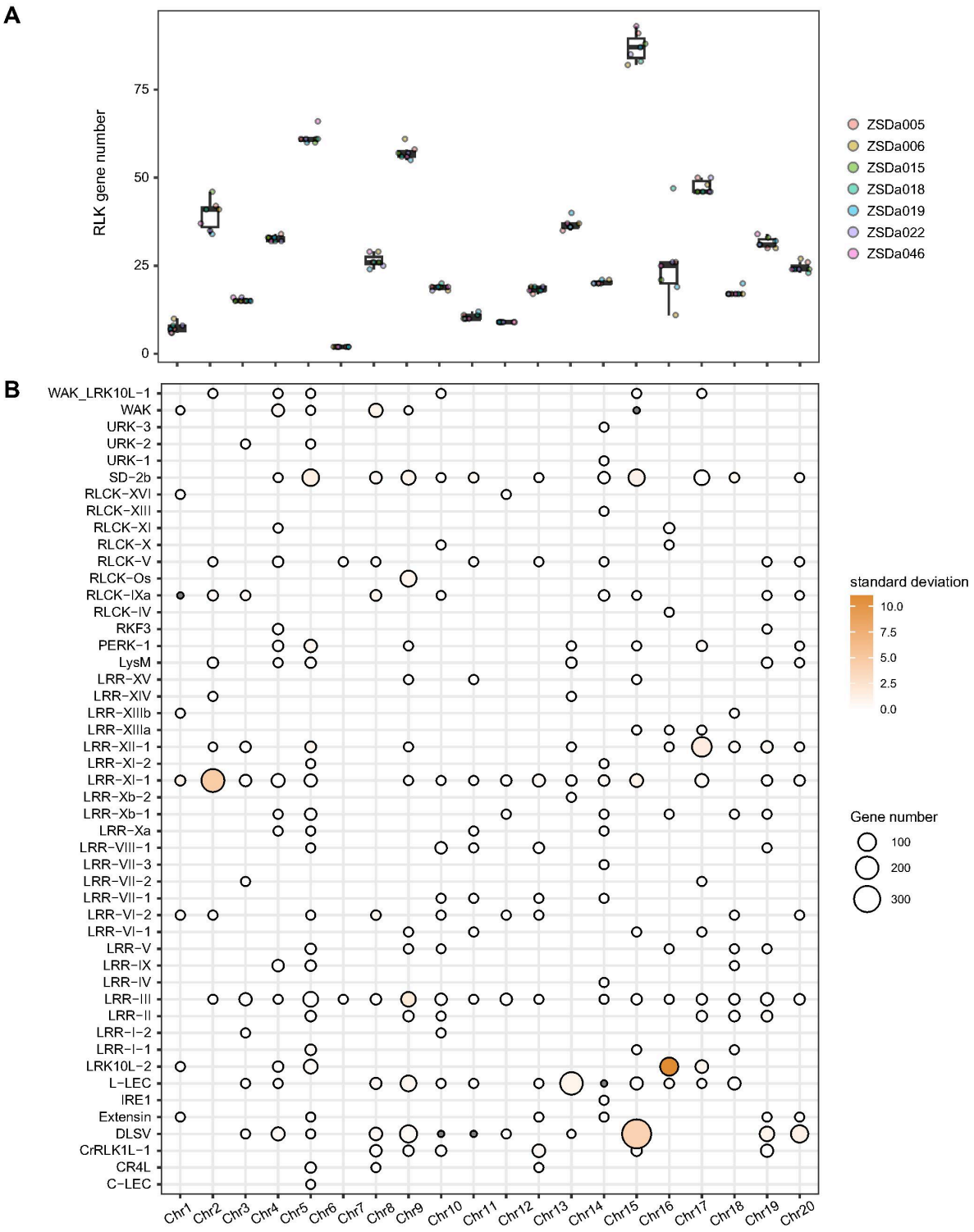

Figure 2

Chromosomal distribution of *D. alata* pan-RLKome. A Distribution of the RLK gene numbers across 20 *D. alata* chromosomes. The RLK gene numbers are presented as boxplots, with points representing each accession in different colours (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, each accession). **B**Distribution of the 46 RLK gene subfamilies across 20 D. alata chromosomes. The RLK gene numbers are presented as a bubble diagram, with the colour and size key shown at the right. Brown point indicates that only one accession possessed RLK gene on corresponding chromosome.

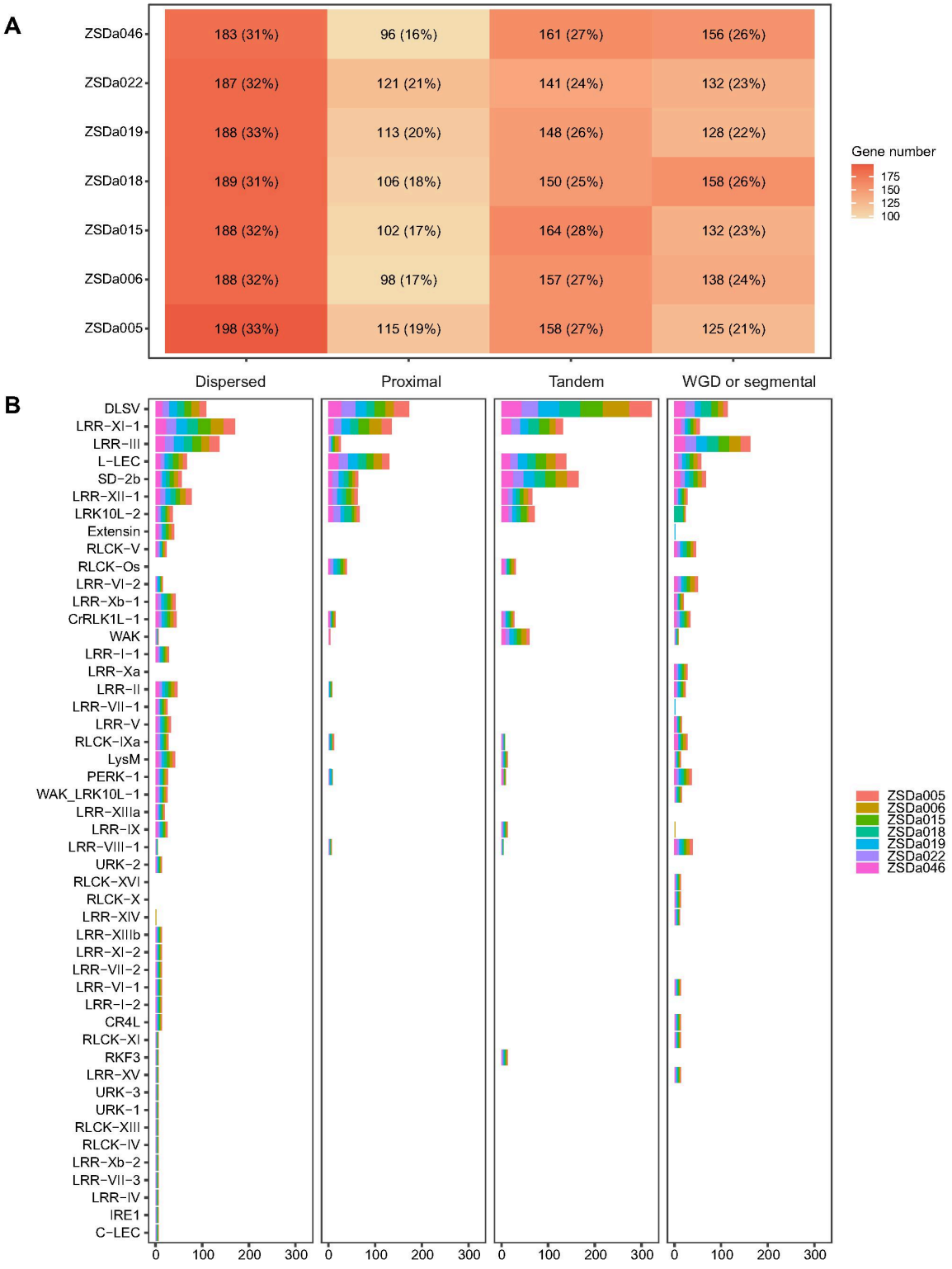

Figure 3

Diverse duplication mechanisms of *D. alata* pan-RLKome. A Number and proportion of the duplication events on RLK genes from seven *D. alata* accessions. The RLK gene numbers are presented as a heat map, with the colour key shown at the right. **B** Distribution of the duplication events on 46 RLK subfamilies from seven *D. alata* accessions. The RLK gene numbers are presented as stacked bar charts. Seven *D. alata* accessions are indicated in different colours.

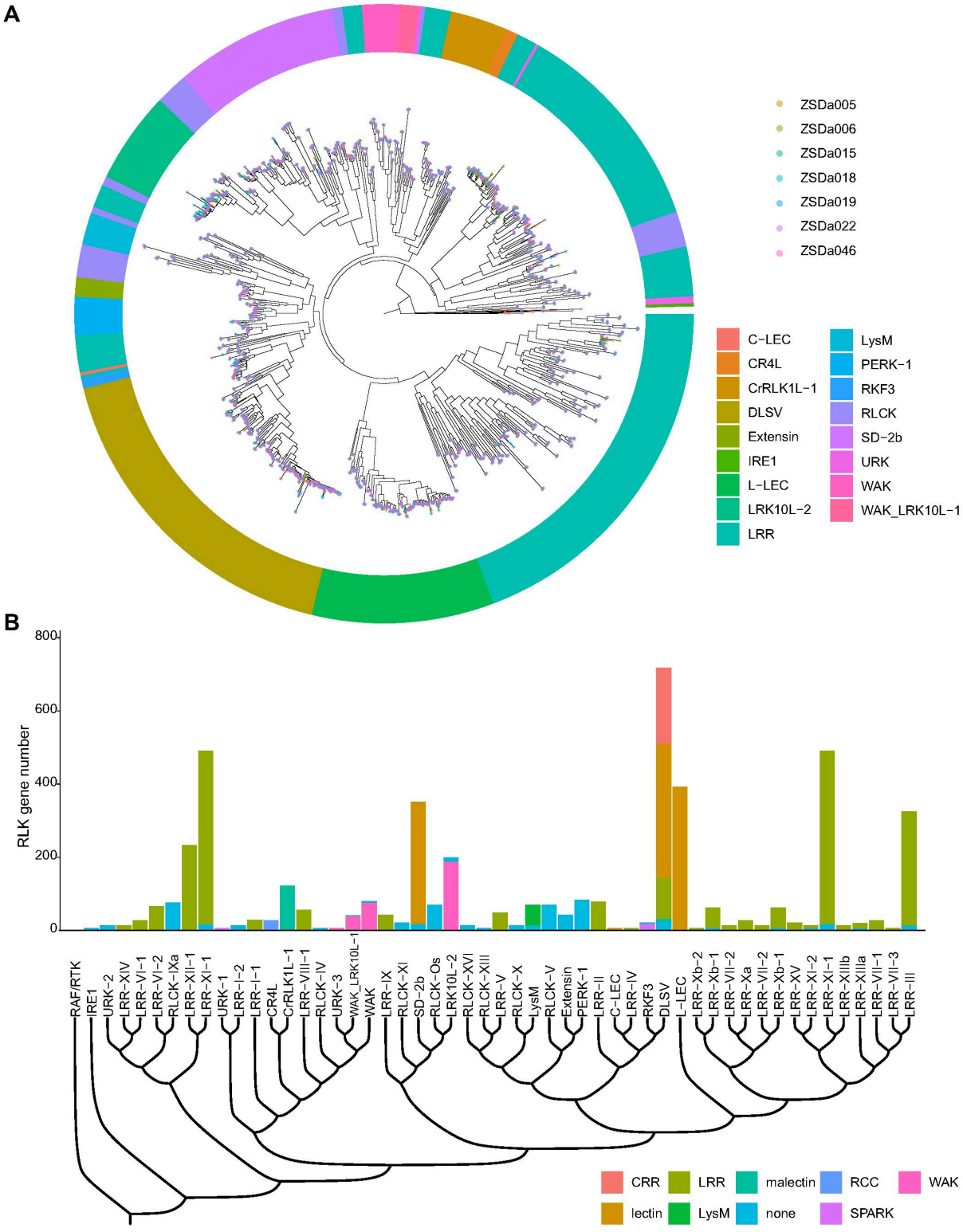

Figure 4

Evolutionary trajectory of RLK gene family in *D. alata*. A Phylogenetic tree of *D. alata* pan-RLKome. Points on the leaf tip represent the accessions in different colours. The outlier ring indicates the RLK subfamilies in different colour. **B** Phylogenetic relationship of RLK subfamilies. The RLK gene numbers are presented as bar plots. RLK with different ECD types are shown in different colours.

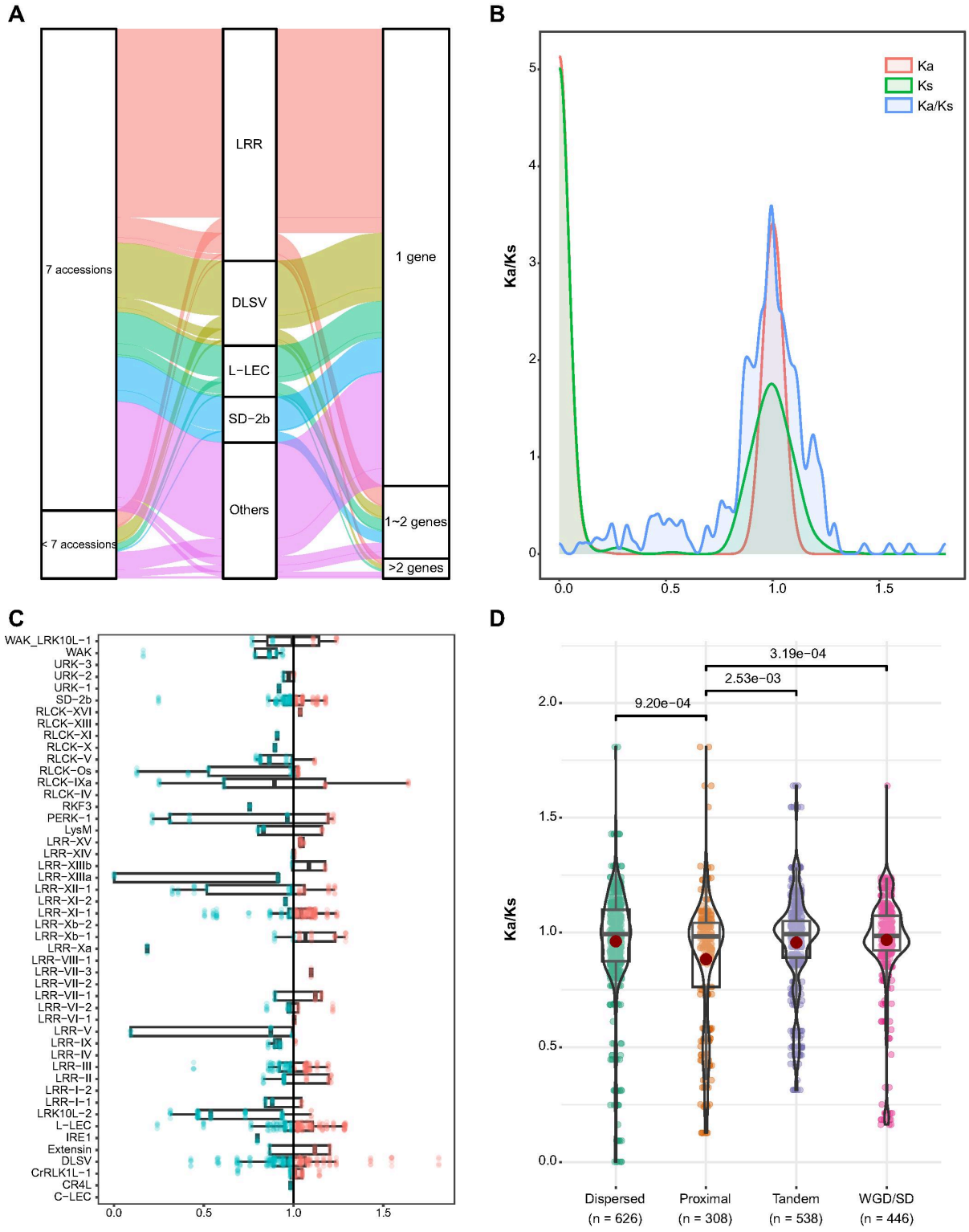

Figure 5

Selection pressure analysis of *D. alata* pan-RLKome. A Composition of *D. alata* RLK OGs. The left section of the Sankey diagram classifies accessions into two groups: “7 accessions” and “< 7 accessions”. The middle section presents different RLK OG subfamilies, including LRR, DLSV, L - LEC, SD - 2b, and others. The right section shows the classification of gene copy numbers per OG: “1 gene”, “1–2 genes”, and “> 2 genes”. **B** Frequency distribution of Ka, Ks, and Ka/Ks of *D. alata* RLK OGs. The values of Ka (red), Ks (green), and Ka/Ks (blue) are presented as a histogram. **C**Distribution of Ka/Ks of *D. alata* RLK OGs in 46 RLK subfamilies. The values of Ka/Ks are presented as boxplots, with points representing each OG (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, each OG). Points with Ka/Ks > 1 are indicated in red, while points with Ka/Ks < 1 are indicated in blue. **D**Distribution of Ka/Ks of D. alata RLK OGs in four duplication types. The values of Ka/Ks are presented as boxplots and violin plots, with points representing each OG (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, each OG). Red dots indicate mean values. Pairwise significance of differences was determined using two-sided Mann-Whitney U-test, and only significant groups are shown.

Supplementary Files

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

- figS18.pdf
- tableS140702.xlsx