Copy Number Variation: A Substrate for Plant Adaptation and Stress Response in Arabidopsis

Piyal Karunarathne[1], Yannick Schäfer[2], Anna Glushkevich[3], Alison D. Scott[3], Leen Abraham[4], Juliette de Meaux[4], Polina Yu Novikova[3], Thomas Wiehe[2], and Laura E. Rose[1]

[1]Institute of Population Genetics, Heinrich-Heine University Düsseldorf, Germany
[2]Institute for Genetics, University of Cologne, Germany
[3]Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany
[4]Institute for Plant Sciences, University of Cologne, Germany

Summary

Copy number variation (CNV) in gene families plays a crucial role in genome evolution and adaptation. However, the evolutionary forces shaping CNV and its ecological relevance remain a key question in evolutionary biology. Here, we investigated CNV across four Brassicaceae species, *Arabidopsis thaliana*, *A. lyrata*, *A. halleri*, and *Arabis alpina*, to identify gene families undergoing rapid expansion or contraction. Using annotated proteomes and a birth–death model of gene family evolution, we identified 231 rapidly evolving gene families, enriched in functions related to regulation, information processing, multifunctional domains, and defense and stress responses.

Focusing on defense- and stress-related gene families, we assessed CNV within populations of *A. thaliana* and *A. lyrata* using long-read assemblies. We identified 29 gene families exhibiting CNV both within and between species. These families showed frequent pseudogenization in roughly two-thirds of cases and functional divergence among paralogs, yet retained signatures of strong purifying selection, highlighting ecological constraint despite structural variation.

Population clustering based on CNV patterns revealed distinct geographic structuring. In *A. thaliana*, populations from Cape Verde, Madeira, and Japan formed divergent clusters, while *A. lyrata* showed separation between North American, Siberian, and European populations, mirroring patterns observed from SNP data.

Ecological modeling further revealed significant associations between CNV and bioclimatic variables, particularly more temperature-derived-variables than precipitation-derived-variables in *A. lyrata*. Additionally, unexplained CNV divergence in *A. thaliana* suggests additional local environmental drivers, such as edaphic conditions or biotic interactions. Our findings demonstrate that CNV in gene families that experienced rapid expansion and contraction among *Arabidopsis* species, contribute to local adaptation across complex environmental gradients.

1 Introduction

Gene duplication and gene loss are fundamental evolutionary processes that generate genetic diversity across genomes and lineages. Among the different forms of structural genomic variation, copy number variation (CNV) represents a major source of heritable genomic variation, defined as insertions, deletions, or duplications of genomic segments ranging from kilobases to

megabases in size (Freeman et al., 2006; Redon et al., 2006). CNVs can encompass entire genes or gene clusters, and influence phenotypic variation through altered gene dosage, expression levels, or regulatory interactions (Zhang et al., 2009; Schrider and Hahn, 2010). In plants, CNVs have been implicated in a wide range of adaptive traits, including disease resistance, flowering time regulation, and responses to abiotic stress (Maron et al., 2013; Cao et al., 2011; Prunier et al., 2017).

The importance of CNVs in plant adaptation is increasingly recognized, particularly in relation to biotic and abiotic stress responses. Several gene families involved in pathogen recognition (e.g. NBS-LRR genes), detoxification, or signaling have been shown to exhibit dynamic CNV patterns both within and between species (Kang et al., 2012; Rizzon et al., 2006; Hanikenne et al., 2008). These gene families can often be described by a birth-and-death model of evolution, whereby gene duplication is followed by functional diversification, neofunctionalization, or pseudogenization, while other paralogs are lost (Nei and Rooney, 2005) (See Otto and Wiehe (2023) for an alternative mechanism: non-homologous cross-over). While the population-level consequences of CNV have been explored in domesticated crops (Muñoz-Amatriaín et al., 2013; Ma et al., 2025), fewer studies have examined its evolutionary role in natural populations across multiple related species.

The Brassicaceae family, which includes the model species *Arabidopsis thaliana* **(L.) Heynh.** and its close relatives, provides an ideal framework to investigate the evolution and ecological consequences of CNV. High-quality reference genomes are available for several Brassicaceae species, and extensive population genomic resources have been generated for *A. thaliana* and *A. lyrata*, enabling fine-scale analyses of structural variation and its ecological correlates (Alonso-Blanco et al., 2016; Hu et al., 2011; Novikova et al., 2016). Moreover, the family exhibits a broad geographic and environmental range, which facilitates testing for associations between CNV and environmental gradients.

The *Arabidopsis* genus provides an exceptional framework for studying gene family evolution and copy number variation due to its well-characterized genomes, diverse life histories, and contrasting demographic backgrounds. *A. thaliana*, a predominantly self-fertilizing annual, has colonized a wide range of habitats across Eurasia and North Africa and exhibits pronounced population structure shaped by glacial history, with genetically distinct relict and post-glacial expansion lineages (Alonso-Blanco et al., 2016; François et al., 2008). In contrast, *A. lyrata* **(L.) O'Kane & Al-Shehbaz** is an outcrossing perennial with a circumboreal distribution, including disjunct populations in Europe, Asia, and North America, providing a valuable comparative system for exploring how different mating systems and range dynamics influence CNV evolution (Hu et al., 2011; Hämälä et al., 2018a). The availability of population-scale whole-genome resequencing, long-read assemblies, and detailed environmental metadata enables high-resolution mapping of CNV and its ecological correlates in both species. Moreover, previous studies in *A. thaliana* have linked CNVs in specific gene families such as NB-LRR disease resistance genes and F-box proteins, to local adaptation and variation in biotic stress response (Todesco et al., 2010; Exposito-Alonso et al., 2019). By leveraging these genomic and ecological resources, our study is uniquely positioned to uncover how gene duplication and loss contribute to adaptive divergence across spatial and environmental gradients in natural populations.

Hence, in this study, we investigate the macro- and micro-evolutionary dynamics of CNVs in gene families of the Brassicaceae. Using genome assemblies from four species and population-scale long-read sequencing data from *A. thaliana* and *A. lyrata*, we identify gene families undergoing rapid expansion or contraction, with a particular focus on those associated with defense and stress response. We further explore population structure and environmental associations based on CNV, to understand the role of gene family evolution in local adaptation. By integrating comparative genomics, population genomics, and ecological modeling, our study sheds light on the evolutionary mechanisms and functional consequences of CNV in plant adaptation.

2 Materials and Methods

2.1 Analysis of gene family expansion and contraction

Gene families were inferred as orthogroups using OrthoFinder2 (Emms and Kelly, 2019), which defines an orthogroup as the complete set of genes descended from a single gene in the last common ancestor of the species considered. This definition aligns with the concept of a gene family as a set of homologous genes that have evolved from a common ancestor through speciation and/or duplication events. By including multiple closely related species and an outgroup, we ensured that orthogroup delineation reflects both recent and ancient gene duplications, improving the resolution of gene family boundaries.

We assessed the rate of gene family evolution in the selected Brassicaceae species using the curated reference proteomes of *A. thaliana*, *A. lyrata*, *A. halleri* **(L.) O'Kane & Al-Shehbaz**, and *Arabis alpina* **L.** (see Tables 3&4 for details). We used *Carica papaya* **L.** as the outgroup, a member of the family Caricaceae (order Brassicales). The data was downloaded from Ensemble reference database (Martin et al., 2022). We only kept the primary proteins (longest transcript of a gene) for all the genes from these proteomes to avoid complications from variations of proteins in the gene family discovery.

In summary, OrthoFinder infers orthogroups using the original OrthoFinder algorithm (Emms and Kelly, 2015) where an all-vs-all BLAST is performed with a bit score to eliminate the bias caused by sequence length. The bit score is then used to group genes (i.e., orthogroups) based on a graph-based clustering approach called MCL (Markov Clustering Algorithm). An unrooted gene tree is inferred for each orthogroup using DendroBLAST. The unrooted species tree is inferred from this set of unrooted orthogroup trees using the STAG algorithm. This STAG species tree is then rooted using the STRIDE algorithm by identifying high-confidence gene duplication events in the complete set of unrooted orthogroup trees. The rooted species tree is used to root the orthogroup trees. Orthologs and gene duplication events are inferred from the rooted orthogroup trees by a novel hybrid algorithm that combines the "species-overlap" method and the duplication-loss-coalescent model. Finally, comparative statistics are calculated.

All orthogroups with high confidence and resolved trees from OrthoFinder were used as gene families for assessing the evolutionary rates of gene families among species. The evolutionary rates based on the gene Birth and Death (BD) model of gene family evolution is calculated using the CAFE V software (Hahn et al., 2005; Mendes et al., 2020). The algorithm in CAFE uses the number of genes in each species for each gene family and a time-calibrated species tree to determine the gene family expansion and contraction rate λ parameter, which represents the rate of gene gain (birth) and loss (death) per gene per million years. CAFE V uses a likelihood ratio test (LRT) to find the best model and Monte Carlo simulations to calculate expected family sizes using the estimated λ parameter. A chi-square test determines the significance.

$$\Delta[L] = 2(logL_1 - logL_0)$$

L_0 - likelihood of the null model ($\lambda = 0$)
L_1 - likelihood of the alternative model ($\lambda > 0$)
$\Delta[L]$ - likelihood ratio test statistic

In summary, CAFE V analyzes changes in gene family size in a way that accounts for phylogenetic history and provides a statistical foundation for evolutionary inferences. The program uses a birth and death process to model gene gain and loss across a user-specified phylogenetic tree. This enables the estimation of an evolutionary rate for gene family evolution as a global (single λ values) or local (multiple λ values) rate. The latter allows for adding more gamma distribution parameters to find the best fit model considering varying levels of evolutionary rates in different gene families.

In order to find the best λ that describes the evolutionary rates among the studied species, we ran CAFE V with gamma parameters varying from $k = 1$ (base model) to $k = 5$ (maximum allowed for five species). The model outputs are log-likelihood and λ estimations for each model. We used both the log-likelihood and AIC/BIC to determine the best model without overfitting the data.

An error model was run to account for genome annotation artifacts and the resulting error factor was used to rerun the base CAFE models. The error factor was 0.023207 and it did not significantly change the λ values of the base model, indicating that the genome annotations are considerably accurate.

We used a maximum p-value of 0.01 to find the significantly fast evolving gene families under the base model. The gene functions of the members of these gene families were predicted as explained below.

2.2 Gene Family Function Assessment and Categorization

We used *A. thaliana* genes to find the predicted functions of the members of the gene families. This was done using custom python and R scripts to search the gene function databases of UniProt (Consortium, 2022) and TAIR (Berardini et al., 2019). Based on the primary gene function, we further categorize the genes into nine broader functional categories; Uncharacterized and uncategorized genes were combined into one category, making it 10 main functional categories.

We further categorized the gene families into two categories based on the functional similarity of the members of the gene families; gene families with all members with the same function were categorized as "conserved" and families with members with different gene functions as "diversified." Gene families with members on the same chromosome within a species were categorized as "co-localized" and those occurring in different chromosomes were categorized as "dispersed." Finally, the gene families were also categorized based on their number of members: "small" – less than five members, "medium" – 5-15 members, "large" – 16 or more members.

2.3 Assessment of Copy Number Variation and Evolutionary Trajectories in Defense- and Stress-Related Gene Families

2.3.1 Selection of Gene Families for CNV Analysis

From the initial set of 231 fast-evolving gene families identified through comparative analysis, we selected 29 gene families that are functionally associated with plant defense and stress response for detailed copy number variation (CNV) analysis within and between species. These families were prioritized due to their ecological relevance and known involvement in adaptive responses in *Arabidopsis* species (e.g., Almaghrabi et al., 2019; Wang et al., 2011). The subsequent analyses are focused on these families.

To assess copy number variation (CNV) within species in selected gene families, we used a curated dataset of long-read *de novo* genome assemblies comprising 137 *Arabidopsis thaliana* assemblies (from NCBI: see Table S3) and 24 *A. lyrata* assemblies (Table S4) generated in the present study and a pangenome study of the species (in prep.). Sequencing details and methods used to generate the *A. lyrata de novo* assemblies are described in the Supplementary Information (**S1**). A complete list of genome assemblies used for CNV detection is provided in Tables S3 & 4. We searched for gene copies in all *de novo* assemblies using nucleotide BLAST (BLAST+) (Camacho et al., 2009a), with reference sequences from all members of the selected gene families. A detailed CNV detection workflow is presented in the Supplementary Information (**S2**).

2.3.2 Classification of Gene Family Copy Number Variation

Gene families were categorized as CNV-present or CNV-absent based on the number of gene copies observed in individual assemblies compared to the number of members in the Col-0 reference genome for *Arabidopsis thaliana* and MN47 reference genome for *A. lyrata*. CNV presence was further classified by the frequency of CNV across sampled individuals: families with CNV in over 75% of assemblies were labeled as *high-frequency CNV*, those with 25–75% as *medium-frequency CNV*, and those below 25% as *low-frequency CNV*. This classification allows evaluation of the stability or variability of gene family size within species.

2.3.3 Population Structuring Based on CNV Profiles

To assess whether CNV patterns reflect population structure or demographic history, we applied multiple correspondence analysis (MCA) to gene family copy number variation across assemblies. MCA is a multivariate ordination technique suitable for summarizing and visualizing patterns in large categorical datasets, where the variables represent discrete states rather than continuous measurements. In this context, each gene family was encoded as a categorical variable in which the different observed copy numbers (0, 1, 2, ...) were treated as distinct states. This representation captures the presence or absence of specific copy states, thereby accommodating the fact that CNV data are inherently discrete and non-linear, and that the absence of a gene (copy number = 0) is itself a biologically meaningful state.

By transforming the data into a reduced set of orthogonal dimensions, MCA preserves the chi-squared distances among samples in the multi-gene-state space, allowing detection of shared CNV patterns across individuals or populations. This is particularly appropriate for CNV datasets because: (i) copy number states do not have a uniform or necessarily linear relationship to functional or evolutionary effects, making continuous ordination methods (e.g., PCA) less suitable; (ii) the method accommodates rare states and zero counts without imposing distributional assumptions; and (iii) the output facilitates interpretation of patterns in terms of categorical co-occurrence (e.g., certain copy number states co-occurring in specific populations). The reduced-dimensional space of MCA was used to visualize population-level clustering, which was then evaluated for correspondence with known geographic distributions or genetic groups, providing insight into whether the CNV structure aligns with historical demography or environmental adaptation.

2.3.4 Ecological Association of Gene Family CNV

To assess environmental drivers of CNV variation in defense- and stress-response gene families, we performed species-specific MCAs for *A. thaliana* and *A. lyrata*, followed by permutation-based environmental fitting. The MCA scores used here represent orthogonal dimensions summarizing individual-level CNV profiles after correction for neutral population structure. Neutral correction was achieved by regressing the MCA coordinates against a covariance matrix derived from putatively neutral loci (four-fold degenerate sites, introns, and intergenic regions) identified from genome-wide SNP data (*A. thaliana*: PacBio long-read; *A. lyrata*: Illumina WGS), following the approach of Gautier (2015) (see Supplementary **S3** for more information on SNP dataset and neutral diversity assessment). This ensured that the residual variation in CNV was less likely to be driven by demographic history or isolation-by-distance.

The residual MCA scores were then analyzed using the `envfit` function from the **vegan** R package (Oksanen et al., 2022) to quantify and test associations with environmental gradients. `envfit` fits each environmental variable to the ordination space by finding the vector in that space that maximizes the correlation between the variable and the sample scores along all MCA dimensions. The significance of this fit is evaluated via permutation: environmental values are randomly shuffled among samples (here, 999 permutations), and the proportion of permutations

with a correlation greater than or equal to the observed correlation yields the p-value. This non-parametric approach makes no assumptions about the distribution of CNV data and can accommodate both continuous and categorical predictors.

Bioclimatic variables (Bio1–Bio19) were obtained from the CHELSA climatic database (Karger et al., 2017) at 30 arc sec resolution (Table S5). To reduce multicollinearity, variables with pairwise Pearson correlation coefficients $r > 0.75$ were sequentially removed prior to fitting. Significant environmental vectors were overlaid on MCA biplots to visualize the direction and strength of their association with CNV variation.

To identify the gene families contributing most strongly to ecological divergence, we calculated the multivariate influence of each gene family in the MCA space. First, we extracted the MCA coordinates for each gene copy number state and computed the squared distance from the origin in the first two MCA dimensions, which capture the majority of variation. These distances reflect the contribution of each CNV state to overall genetic differentiation. We then aggregated these contributions across all copy number states belonging to the same gene family. The gene families were subsequently ranked based on their cumulative contribution, and the top five most influential families were identified as key candidates underlying ecological differentiation in copy number variation.

2.3.5 Inference of Evolutionary Trajectories

To investigate the evolutionary dynamics of the selected gene families, we computed molecular evolutionary statistics indicative of selection pressure. These included average nucleotide diversity (π), Watterson's theta (θ_W), and the nonsynonymous to synonymous nucleotide diversity ratio π_A/π_S, and Tajima's D. Together, these metrics reveal whether gene families are evolving under purifying, neutral, or positive selection. Patterns of divergence and diversity were compared across gene families to identify signals of functional innovation (e.g., subfunctionalization or neofunctionalization) or pseudogenization.

3 Results

3.1 Comprehensive Identification of Gene Families Using OrthoFinder2

Our gene family assignment using OrthoFinder2 yielded a well-resolved, rooted species tree with strong bootstrap support (Fig. 1), consistent with established phylogenies of the order *Brassicales* (e.g., Hendriks et al., 2023). This supports the reliability of the phylogenetic inference provided by OrthoFinder2.

Across all five species, OrthoFinder2 annotated a total of 136,809 genes, of which 92% were assigned to 22,083 orthogroups. Among the in-group taxa, over 93% of genes were clustered into orthogroups, whereas the outgroup had a lower assignment rate of 79% (Table 1). *Arabidopsis halleri* showed the highest proportion of genes in orthogroups (97.5%), while *A. alpina* exhibited the greatest number of species-specific orthogroups (292, comprising 10% of all its genes). In contrast, *A. thaliana* had the fewest species-specific orthogroups (99, 1.2%).

Subsequent results focus exclusively on the in-group species. A large proportion of genes (46-66%) were found in single-copy orthogroups, with *A. thaliana* having the highest percentage (66.7%). *A. halleri* showed the greatest number of large gene families (129 orthogroups with 16 or more genes), while *A. thaliana* had the fewest (48 orthogroups).

OrthoFinder also predicted duplication events, with the highest number occurring in the last common ancestor of *A. thaliana*, *A. lyrata*, and *A. halleri* (2,857 events; node N2 in Fig. 1). A considerable number of duplications (2,392) were also inferred for the ancestor of *A. alpina* and the Arabidopsis lineage. Notably, *A. halleri* experienced a substantial number of duplications at the terminal branch, indicating lineage-specific gene family expansion.

Table 1: Summary statistics of gene content and orthogroup (OG) assignment across five species used in the OrthoFinder2 analysis. The table reports the total number of predicted genes, the proportion assigned to orthogroups (shared gene families), and the number and percentage of unassigned genes for each species. It also shows how many orthogroups each species contributes to ("OGs containing species") as well as the number and proportion of species-specific orthogroups (i.e., orthogroups unique to a single species).

	A. alpina	***A. halleri***	***A. lyrata***	***A. thaliana***	***C. papaya***
Genes	21,609	28,722	31,073	27,654	27,751
Genes in OGs	20,144 (93.2%)	27,999 (97.5%)	29,383 (94.6%)	26,442 (95.6%)	21,896 (78.9%)
Unassigned genes	1,465 (6.8%)	723 (2.5%)	1,690 (5.4%)	1,212 (4.4%)	5,855 (21.1%)
OGs containing species	13,443 (60.9%)	18,734 (84.8%)	19,735 (89.4%)	19,228 (87.1%)	13,315 (60.3%)
Species-specific OGs	292	241	321	99	889
Genes in species-specific OGs	2,162 (10.0%)	1,505 (5.2%)	992 (3.2%)	341 (1.2%)	4,635 (16.7%)

3.2 Modeling Gene Family Evolution Using CAFE V

We analyzed the evolutionary dynamics of the 22,083 gene families using CAFE V, which models gene family size changes under a birth-and-death process. The input included the gene copy number per species and the dated, rooted species tree. The species tree from OrthoFinder was calibrated using the `makeChronosCalib` function from the `APE` R package (Paradis and Schliep, 2019), based on a divergence estimate of 4-6 Mya between *A. thaliana* and the rest of the *Arabidopsis* species (Novikova et al., 2016; Hohmann et al., 2015) .

To identify the most suitable evolutionary rate model, we compared models with one to five gamma distribution of evolutionary rate classes ($\gamma = 1$–5). Both Akaike (AIC) / Bayesian (BIC) information criteria and log-likelihood supported the base model ($\gamma = 1$) as the best fit (Fig. S1). The estimated average gene family turnover rate under this model was $\lambda = 0.0118$, with the maximum rate for this topology being $\lambda_{\max} = 0.0238$.

Incorporating an error model to account for annotation errors did not significantly alter the estimated rate ($\lambda = 0.0109$, [error rate] $\varepsilon = 0.0232$), confirming the robustness of the base model. We therefore retained the base model for downstream analyses.

3.3 Detection of Rapidly Evolving Gene Families

Based on the base model, CAFE V identified 1,592 gene families with significant expansion or contraction ($p<0.05$). For a more stringent and biologically relevant subset, we applied a significance threshold of $p<0.01$, resulting in 821 gene families classified as "fast-evolving".

To facilitate comparative and functional analyses, we filtered this list further by retaining only the gene families present in at least three species and containing at least one gene from *A. thaliana*, enabling downstream annotation and functional prediction. This yielded a final set of 231 fast-evolving gene families for in-depth exploration.

Among these, 79 were small (fewer than five genes), 112 were medium-sized (5–15 genes), and 40 were large (more than 15 genes), highlighting the size diversity of gene families undergoing rapid evolutionary changes among the studied lineages.

3.4 Functional Categories of Fast-Evolving Gene Families

Based on *A. thaliana* gene ontology (GO) predictions, the 231 significantly fast-evolving gene families were classified into 10 major functional categories (Fig. 2). The largest proportion comprised **genetic information & regulation** (23.4%), followed by **multi-function proteins** (22.1%). **Defense and stress response genes** accounted for 12.6%, and 15.2% were uncategorized. Gene families related to **structural dynamics** and **photosynsthesis & energy related proteins** were the least represented, constituting only 1.3% each.

Notably, many gene families contained members with diverse gene ontology terms. In such cases, we assigned the family to the most frequently occurring functional category among its

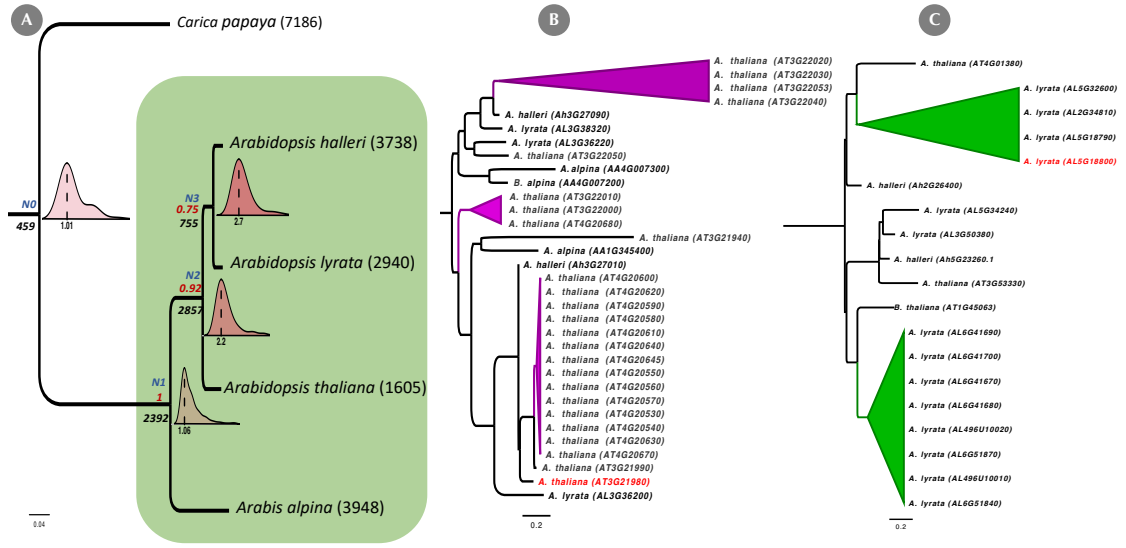

Figure 1: **Phylogenetic trees inferred from OrthoFinder2 analysis**. **A. Species tree**; Brassicaceae (in-group) species are highlighted green. **Blue font:** shows the node number corresponding to the gene duplication assessment by OrthoFinder2, **Red font:** shows bootstrap support values for each branch, **Black font:** shows the number of gene duplication events at the respective most recent common ancestors, **Density plot:** shows the λ value (gene family evolutionary rate) distribution across all gene families that showed significant evolutionary rate obtained from CAFE V analysis at the respective node; the intensity of the color represents the mean λ, which is also shown by the dashed line. Number of duplicated genes in each species is shown within brackets. **B&C. Resolved Gene Trees** of two significantly fast-evolving gene families showing withing species gene family expansion in (B) *A. thaliana* (OG0000093) and (C) *A. lyrata* (OG0000347); Most frequently pseudogenized genes are indicated in red.

members. A comprehensive list of gene families, their constituent genes, and assigned functions is provided in Table S11. Further, a GO enrichment analysis indicated that both "Defense and Stress-Response" and "Biosynthesis and Metabolism" related gene families were significantly enriched with former having the highest enrichment (gene ratio: 0.05; adjusted p-value: 1.8×10^{-29}: see Fig 2C and Table S6).

3.5 Functional divergence, pseudogenization, and chromosomal organization of gene families

Our analysis of gene family functional divergence revealed that more than 50% (124) of the gene families showed evidence of functional diversification, meaning that different members within a family code for functionally distinct proteins. These functionally diverging families were predominantly medium- and large-sized gene families. Notably, nearly all gene families whose members encoded multiple protein domains exhibited functional divergence (i.e., members do not code the same multiple protein domains).

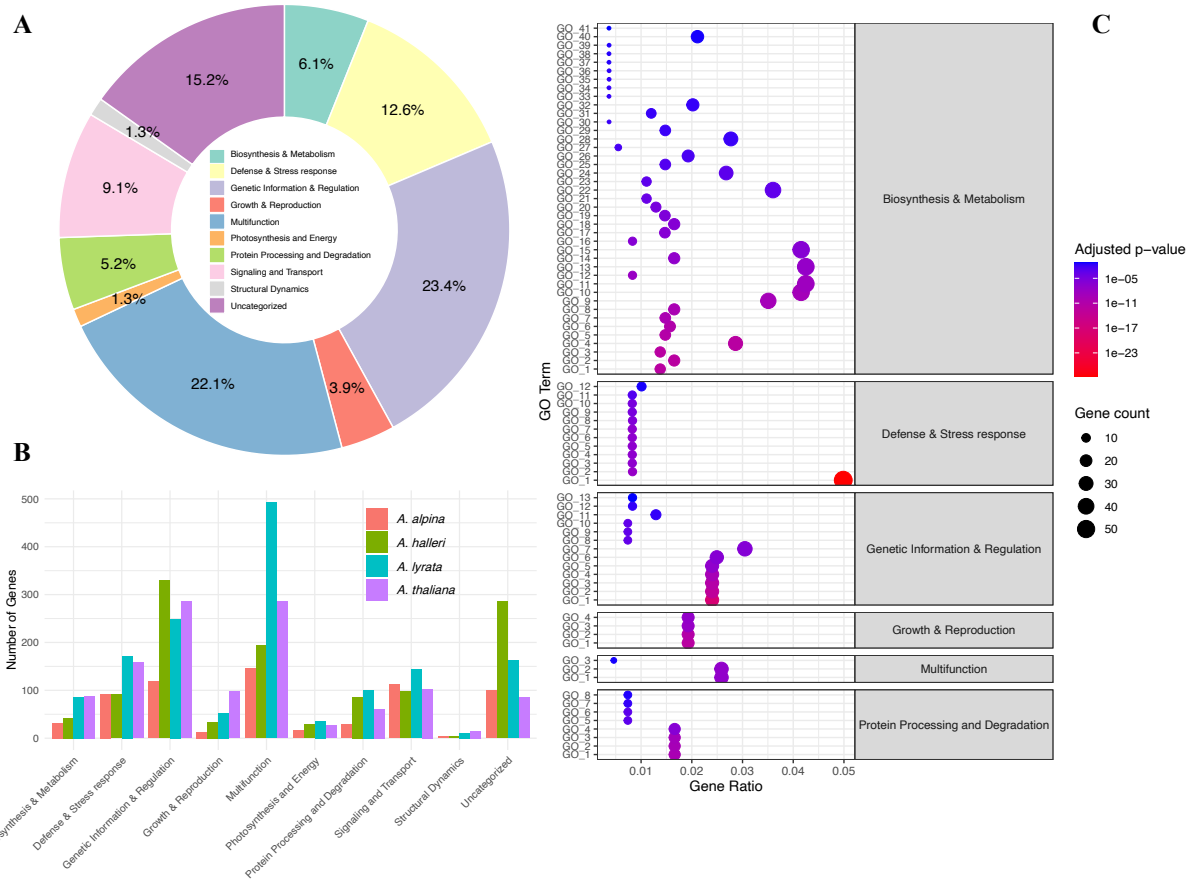

Figure 2: **Functional classification and GO enrichment of fast-evolving gene families across *Arabidopsis* species. (A)** Pie chart showing the proportion of functional categories among rapidly evolving gene families. Functional annotations were assigned using *A. thaliana* gene models, with Gene Ontology (GO) data retrieved from The Arabidopsis Information Resource (TAIR) and UniProt databases. Each category represents the dominant annotated function for at least one gene within a rapidly evolving orthogroup. **(B)** Bar plot showing the number of genes assigned to each functional category across the four species analyzed, illustrating interspecific variation in the expansion of functionally defined gene families. **(C)** GO enrichment analysis of the functional categories. GO terms were retrieved from TAIR, grouped by gene function, and filtered to retain only significantly enriched categories (adjusted p-value after Benjamini–Hochberg correction). GO terms are labeled arbitrarily (GO-1 to GO-n) for clarity; a complete list of GO terms and enrichment statistics is provided in Table S6.

We also assessed the chromosomal distribution of gene family members, focusing on whether members were **co-localized** on the same chromosome or **dispersed** across different chromosomes. This organization can reflect the evolutionary origin of gene families (e.g., local or tandem duplication versus translocations) and reveal genome structural rearrangements. In all species examined, except *Arabidopsis lyrata*, about one-third of gene families ($\sim$ 45) had members co-localized on the same chromosome (Table S1). In contrast, in *A. lyrata*, only 15% of gene families showed co-localization, suggesting that different mechanisms of gene family expansion, possibly influenced by transposable elements (TEs: see TE analysis below), or other chromosomal rearrangements have shaped its genome structure.

3.5.1 Pseudogenization and TE association of Defense and Stress-Response gene families

To investigate evolutionary signatures and potential functional shifts in rapidly evolving gene families, we focused on 29 families associated with "Defense and Stress-Response" functions. Further analyses were centered on these families.

We evaluated the frequency of pseudogenization among gene family members that underwent recent duplication-driven expansions or contractions. Gene members extracted from long-read *de novo* assemblies of *A. thaliana* (137 accessions from NCBI) and *A. lyrata* (23 assemblies from this study) were reannotated. BLAST searches (BLAST+ Camacho et al., 2009b) identified sequences with ≥90% identity to reference genes. These sequences were reannotated using HELIXER (version 0.3.5: Stiehler et al., 2020) and compared to the corresponding reference annotations. A gene was classified as a putative pseudogene if it contained an early stop codon (i.e., presence of a stop codon withing the range of 26-74% of the length of the complete gene) or if over 30% of its gene sequence was missing compared to the reference. Genes with over 10% amino acid divergence relative to the reference were classified as putatively neo- or sub-functionalized. According to this criteria, among the 29 gene families examined, over 70% contained at least one pseudogene in both species (Table S2). All families exhibiting pseudogenization had undergone recent duplication events (see Fig. 1 B&C and Figs. S5-34). In contrast, only 12% of gene families had at least one gene predicted to be neo- or sub-functionalized.

Finally, we assessed the extent of transposable element (TE) association with the rapidly evolving defense- and stress-response gene families by running `RepeatMasker` on long-read de novo assemblies of both species (see Supplementary S4 for details). The presence of known TEs that are associated with gene copy number variation (Morgante et al., 2005; Kapitonov and Jurka, 2007) was evaluated within 1,000 base pairs upstream or downstream of each gene, as well as within gene bodies. In *A. thaliana*, 58.8% (93 out of 158) of the assessed genes were associated with at least one TE in at least one assembly. In contrast, *A. lyrata* showed extensive TE association, with 90% (155 out of 172) of genes linked to at least one TE. In order to compare the enrichment of TEs in multi-copy rapidly evolving gene families, we also assessed 50 gene families with only one copy in all studies species. Both species showed significant enrichment of TEs in rapidly evolving defense and stress response gene families compared to single copy genes (Wilcoxon rank sum test p-value $<<$ 0.001). On average *A. thaliana* has 0.32 TEs per gene while single copy genes have 0.0016 TEs per gene. Similarly, in *A. lyrata*, defense and stress response gene families have 0.46 TEs per gene while single copy genes show an average of 0.281, indicating that *A. lyrata* harbors more TE in its genome (see Fig. 3 boxplots). A gene family was considered TE-associated if at least one of its members exhibited TE proximity. Based on this criterion, 22 of *A. thaliana* gene families and 27 of *A. lyrata* gene families were classified as TE-associated (Fig. 3).

These findings suggest that TE activity may play a substantial role in shaping the genomic architecture and evolutionary dynamics of stress-related gene families, particularly in *A. lyrata*. The higher frequency of TE–gene associations in *A. lyrata* coincides with its greater gene family expansion rates and reduced chromosomal co-localization, implying that TE-mediated duplication and rearrangement events could underlie both phenomena. Moreover, the strong correlation between TE proximity and pseudogenization suggests that TEs may not only facilitate gene duplication but also contribute to gene decay, potentially generating raw material for evolutionary innovation. This highlights the dual role of TEs as both drivers of genomic instability and facilitators of functional divergence in plant genomes.

3.6 Copy number variation within and between species

3.6.1 Gene family size variation within species

To investigate copy number variation (CNV), variation in the number of gene family members, within species, we focused on genes related to "Defense and Stress Response." These families were selected not only because of their ecological importance but also to explore the evolutionary dynamics of CNV in gene families under strong selective pressures. Gene sequences were extracted from *de novo* assemblies of *A. thaliana* and *A. lyrata* as described above.

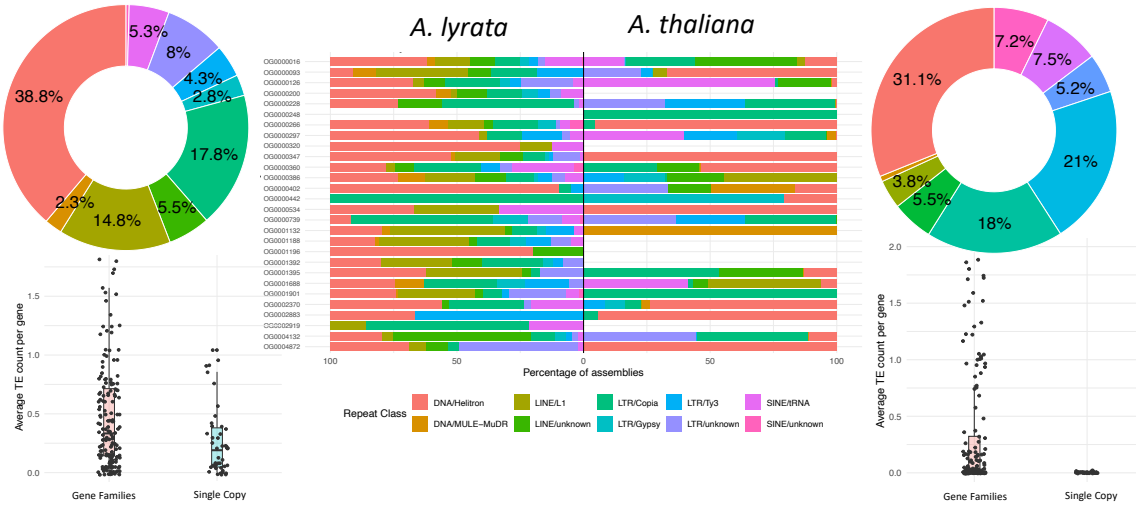

Figure 3: **Summary of TE enrichment analysis with rapidly evolving defense and stress response gene families.** The barplots show the proportional composition of TE classes for gene families across all studied individuals. Gene family labels are on the y-axis and pecentages are on the x-axis. The doughnut plots show the proportion of TE-classes associated with all genes. Boxplots show the average number of TEs per gene present in defense and stress response gene family genes (Gene Families) vs. single copy genes (Single copy), with jitter showing all data points. A Wilcoxon rank sum test showed significant enrichment of TEs (p-value $<<$ 0.001) in defense and stress-response related gene families than in single copy genes in both species.

Among the defense- and stress-related gene families examined, 68% (20 families) in *A. thaliana* and 85% (24 families) in *A. lyrata* exhibited CNV. Moreover, three families in *A. thaliana* and five families in *A. lyrata* showed CNV in nearly all analyzed assemblies (Table 1; Fig. S2 & 3). Families that did not show CNV typically contained fewer than three members in the reference genome. Although we initially used the number of gene copies in the reference genomes as the baseline expectation, recalculating CNV frequencies based on the median gene number across assemblies yielded similar results. This indicates that while small gene families can evolve rapidly between species, they tend to remain relatively stable within species.

3.6.2 Geographic structuring of copy number variation

We next examined whether patterns of CNV in defense- and stress-response gene families exhibited geographic structure within *A. thaliana* and *A. lyrata*. To do this, we performed multiple correspondence analysis (MCA), the categorical analogue of principal component analysis (PCA), on CNV profiles.

In both species, CNV patterns showed clear geographic separation of populations. In *A. thaliana*, populations from Cape Verde, Madeira, and Japan formed distinct clusters, well separated from European populations. Relict European populations also formed separate clusters, with a gradient from relict to non-relict groups (Fig. 4). Similarly, in *A. lyrata*, North American and Siberian populations clustered apart from European populations, reflecting geographic patterns previously observed in SNP-based studies.

In *A. lyrata*, the first dimension of the MCA captured the divergence of North American populations from the rest. After removing North American samples and reanalyzing the European and Siberian populations, the MCA revealed finer geographic structure: eastern Siberian populations separated from European populations, and eastern European populations clustered closer to western Siberian populations than to western European ones (Fig. 4). These results highlight the strong geographic signature underlying CNV in defense- and stress-response gene families.

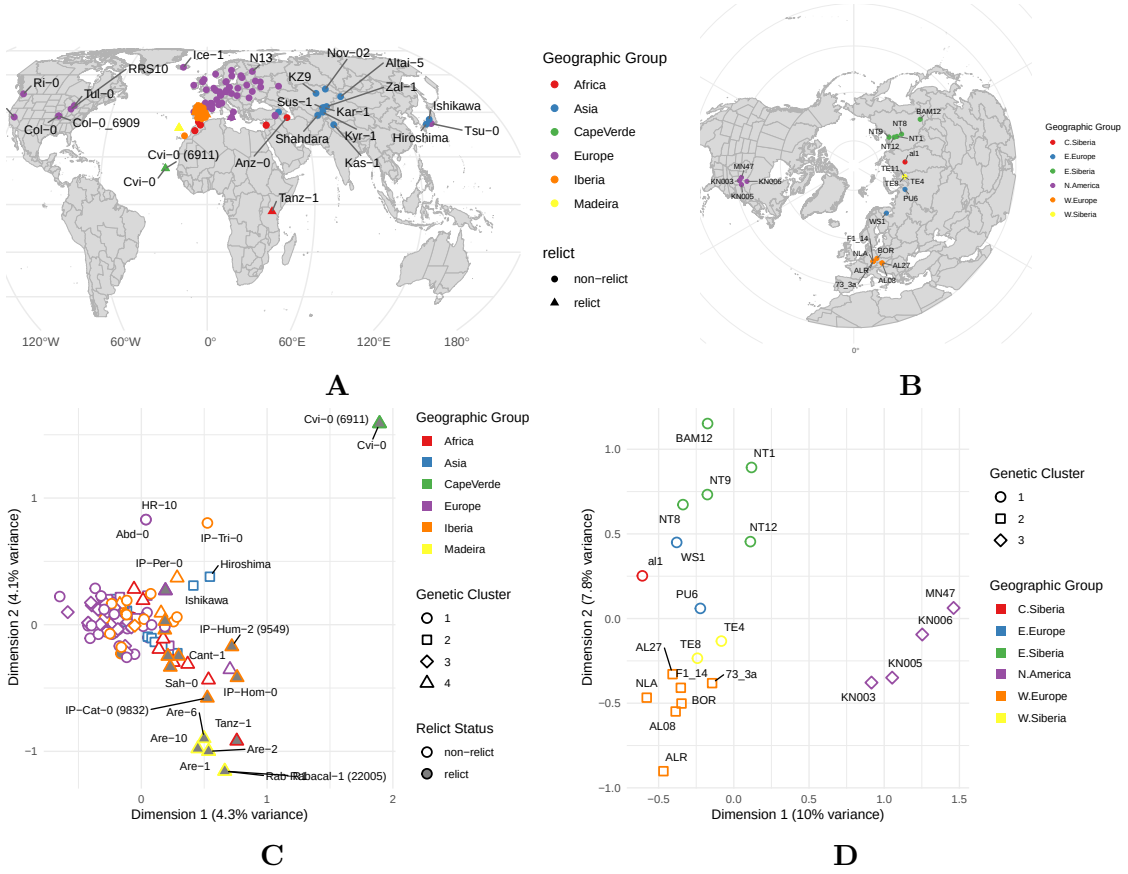

Figure 4: **Geographic structuring of copy number variation in defense- and stress-response gene families. A:** Population locations of *A. thaliana*; colors and shapes fillings (i.e., grey and white) represent genetic clusters according to (Wlodzimierz et al., 2023; Lian et al., 2024; Alonso-Blanco et al., 2016). **B:** Population locations of *A. lyrata*; colors represent genetic clustering according to (Scott et al., 2024). **C:** MCA plot for *A. thaliana* populations. **D:** MCA plot for *A. lyrata* populations. The plot character shapes in C and D represent the genetic clusters determined in this study using putatively neutral SNPs of 136 long read sequences of *A. thaliana* and WGS of *A. lyrata* (see Fig. S4). Map border delineations correspond to the TDWG level III floristic countries according to Brummitt et al. (2001).

3.7 Copy Number Variation and Ecological Divergence

To investigate the relationship between copy number variation (CNV) and environmental variables, we performed ecological association analyses on gene families related to defense and stress responses in *Arabidopsis thaliana* and *A. lyrata*. Our results revealed significant associations between gene copy number and specific bioclimatic variables, suggesting a potential role for CNV in local adaptation in both species.

In *A. thaliana*, multiple climatic variables showed strong associations with CNV patterns. These included temperature-related variables such as Bio3 (Isothermality), Bio4 (Temperature seasonality), Bio8 (Mean temperature of the wettest quarter) and Bio10 (Mean temperature of the warmest quarter), as well as precipitation-related variables including Bio13 (Precipitation of the wettest month) and Bio14 (Precipitation of the driest month), all with $p < 0.001$ (see Fig. 5). After correcting for population structure using covariance matrix derived from genome-wide variation, only one precipitation-related variable (Bio13) and all temperature related variables remained statistically significant ($p < 0.001$), indicating that CNV patterns are decoupled from neutral population structure. While southern populations, except for the Cape Verde Islands, showed a strong association with precipitation variables, northern populations were more strongly associated with temperature variables. This suggests that CNV in *A.*

thaliana is influenced by both temperature and precipitation gradients, with potential ecological implications for local adaptation to varying climatic conditions.

In contrast, in *A. lyrata*, CNV patterns were significantly associated with three temperature-related variables: Bio3 (Isothermality), Bio5 (Max Temperature of Warmest Month) and Bio7 (Temperature annual range) with $p < 0.001$ and one precipitation related variable Bio15 (Precipitation Seasonality)(Fig. 5). This association remained robust after controlling for population structure ($p < 0.01$), suggesting that temperature variability is a primary ecological driver of CNV in this species. Interestingly, Siberian populations were separated from both European and North American populations in the MCA space, indicating that CNV patterns in *A. lyrata* are strongly influenced by geographic and climatic factors, particularly temperature gradients. This suggests that CNV in *A. lyrata* may reflect local adaptation in CNV to specific environmental conditions, despite the strong population genetic structure observed in this species.

Taken together, these findings indicate that CNV in defense and stress-related genes is locally structured by environmental gradients, particularly temperature, in both *A. thaliana* and *A. lyrata*. In *A. lyrata*, the congruence between CNV patterns and precipitation variables suggests ecological adaptation despite strong population genetic structure. In *A. thaliana*, while relic and non-relic lineages exhibit CNV differences partially explained by temperature, the strong isolation of peripheral populations such as those from the Cape Verde Islands and Madeira may reflect the influence of unmeasured local environmental variables. These could include soil and soil related factors, topographic heterogeneity, or biotic pressures, all of which merit further investigation in future studies.

We used squared distances in the MCA space to identify the gene families contributing most strongly to ecological divergence. The top three families with the highest squared distances were *OG0000442* (encodes a plant thionin family protein), *OG0000093* (cysteine-rich receptor-like kinase), and *OG0000266* (Intracellular immune receptor: Encodes a member of the WRKY Transcription Factor (Group II-e) family) in *A. thaliana*, and *OG0000016* (ecodes Tetratricopeptide repeat (TPR)-like superfamily protein), *OG0001132* (Disease resistance protein (NBS-LRR class) family), and *OG0000347* (plastocyanin-like domain-containing protein/copper ion binding / electron carrier protein) in *A. lyrata* (see Tables S9.1 & S9.2). These families are all involved in pathogen recognition, disease resistance, and response to abiotic stress highlighting their potential role in ecological adaptation to biotic and abiotic stressors. The strong association of these gene families with ecological divergence underscores the importance of disease resistance mechanisms in shaping local adaptation in both species. The consistent presence of disease resistance genes among the top contributors suggests that CNV in these families may provide a selective advantage in varying environmental conditions, particularly in response to pathogen pressures. Similarly, the involvement of genes related to abiotic stress responses in the top contributors indicates that CNV may also play a crucial role in adaptation to changing climatic conditions, such as temperature and precipitation variability.

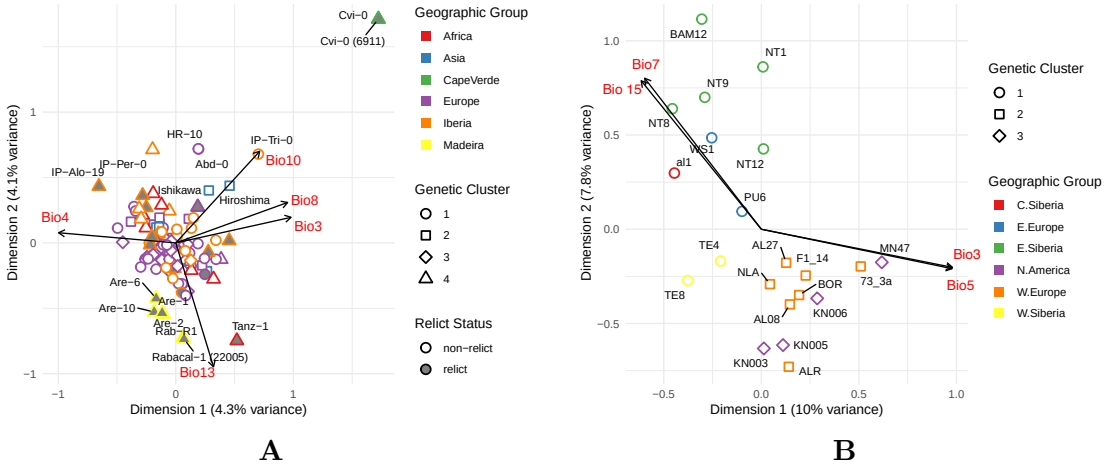

Figure 5: **Environmental MCA (multiple correspondence analysis) of ecological association of CNV in defense- and stress-response gene families. A:** *A. thaliana*. **B:** *A. lyrata*. Colors and plot characters are as in Figure 4. The arrows represent the direction and strength of association with environmental variables, with longer arrows indicating stronger associations. Both geographical clusters and genetic clusters are indicated by plot character colors and shape; additionally for *A. thaliana*, the relict status as in (Alonso-Blanco et al., 2016) is also indicated by grey filling of the plot character. Genetic clusters of populations of both species were determined using hierachical clustering of covariance matrix derived from putatively neutral loci (four-fold degenerate sites, introns, and intergenic regions) from long-read sequences of the two species (Tables S10.1 & S10.2). Bioclimatic variables (i.e., Bio...) are from CHELSA database (Karger et al., 2017) (see table S5 for details).

3.8 Selection Patterns and Functional Divergence in Defense-Related Genes

We assessed evolutionary pressures acting on defense- and stress-response-related gene families in *A. thaliana* and *A. lyrata* using neutrality tests and substitution rate analyses. Observed nucleotide diversity (π) across gene families typically ranged from 1–3% in *A. thaliana* and 1–7% in *A. lyrata*, aligning with prior estimates (Fig. 6). However, a subset of large and medium-sized families exhibited elevated diversity levels, with $\pi > 10\%$ in both species. Notably, these families were also enriched in pseudogenes, suggesting relaxed selection or redundancy due to recent duplications.

Synonymous and nonsynonymous polymorphism (π_S and π_A) ranged between 0.5–1% in both species, although *A. lyrata* showed greater inter-family variation (Fig. 6 and Tables S8.1 & S8.2), consistent with its higher overall nucleotide diversity and population structure.

Tajima's D values for most genes fell between 0 and –1.5, indicating prevalent purifying selection and an excess of low-frequency polymorphisms (see Table S7.1 & 7.2). Only a small proportion of genes had D-values close to zero, suggestive of neutral evolution. Interestingly, a distinct subset of genes exhibited D-values greater than +1.5, pointing to strong positive selection and potentially recent adaptive evolution.

3.8.1 Gene Family Expansion Facilitates Functional Innovation

Genes under strong positive or neutral selection were almost exclusively found in families containing more than 10 gene copies in both species. These multi-copy families also showed higher rates of predicted sub- and neo-functionalization, with individual members diverging in sequence or domain composition relative to their family consensus. This suggests that gene duplication facilitates evolutionary experimentation, where some gene copies retain core functions under purifying selection, while others explore new functional roles under relaxed constraints or directional selection.

Additionally, these gene families often showed greater structural plasticity, manifested as gene relocations across chromosomes, and a higher incidence of pseudogenization. This combination of relaxed constraint on some paralogs and selection-driven divergence in others supports

a model where copy number expansion provides both evolutionary robustness and opportunity for innovation. Together, these findings highlight how gene family expansion not only buffers core functions but also creates substrates for adaptive responses to biotic and abiotic stress.

4 Discussion

4.1 Gene Family Evolution Reflects Phylogenetic Divergence and Life History

The overall patterns of gene family evolution observed in our analysis closely mirror established phylogenetic relationships among Arabidopsis species, consistent with expectations that gene duplication and loss accumulate along species divergence (Lynch and Conery, 2000; Hahn et al., 2005). The orthogroup inference performed with OrthoFinder2 strongly supports the known phylogeny of the genus Arabidopsis, including the close relationship between *A. halleri* and *A. lyrata* relative to *A. thaliana* (Novikova et al., 2016; Murat et al., 2015), validating the accuracy of the gene family assignments (see Fig. 1).

Notably, we observe elevated gene duplication rates in *A. halleri* and *A. lyrata*, which may reflect species-specific life history traits and adaptive processes (see Table 1). In *A. halleri*, increased gene copy numbers have been repeatedly linked to adaptation to metal-rich soils and metal hyperaccumulation (Hanikenne et al., 2008), particularly for genes involved in metal transport and homeostasis. Similarly, *A. lyrata* occupies a wide range of climatic zones and ecological gradients across its geographic distribution, from boreal to temperate environments, potentially driving gene duplication as a mechanism to facilitate local adaptation (Ross-Ibarra et al., 2008). Gene duplication may provide raw material for functional diversification, allowing populations to adapt to heterogeneous environmental conditions (Flagel and Wendel, 2009).

In contrast, *A. thaliana* exhibits overall lower gene duplication rates. This pattern may partially reflect its predominantly selfing mating system, which reduces effective population size and limits the efficacy of selection in maintaining duplicated gene copies (Wang et al., 2012). Additionally, smaller effective population size increases the probability of gene loss through genetic drift, contributing to the observed reduction in gene copy number variation (Panchy et al., 2016).

4.2 Birth-Death Dynamics Are Dominated by Steady Turnover

Our gene family evolution analysis reveals that most gene families in *Arabidopsis* genus exhibit patterns consistent with homogeneous turnover dynamics. The support for the base model assuming a single turnover rate (λ) across all gene families suggests that, for the majority of gene families, gene duplication and loss events occur under a uniform birth-death process without substantial lineage- or family-specific shifts. This observation aligns with previous findings that most gene families across diverse taxa evolve under relatively steady turnover rates over macroevolutionary timescales (Han et al., 2013; Demuth and Hahn, 2009).

The estimated turnover rate in our study ($\lambda \approx 0.0118$) is comparable to rates reported in earlier studies of plant and animal lineages. For example, studies in mammals have estimated turnover rates ranging from approximately 0.001 to 0.01 per gene per million years (Demuth et al., 2006), while estimates in flowering plants have been reported in a similar range, with λ values typically between 0.005 and 0.02 depending on the taxonomic group and model assumptions (Cannon et al., 2004; Panchy et al., 2016). These comparable rates suggest that gene family turnover in Arabidopsis occurs at rates broadly similar to those observed across angiosperms and other multicellular eukaryotes.

Consistent with these turnover dynamics, the majority of gene families in our dataset are not experiencing extreme expansions or contractions. Instead, they appear to be evolving under a steady process of gene gain and loss, in which occasional duplication or loss events occur but

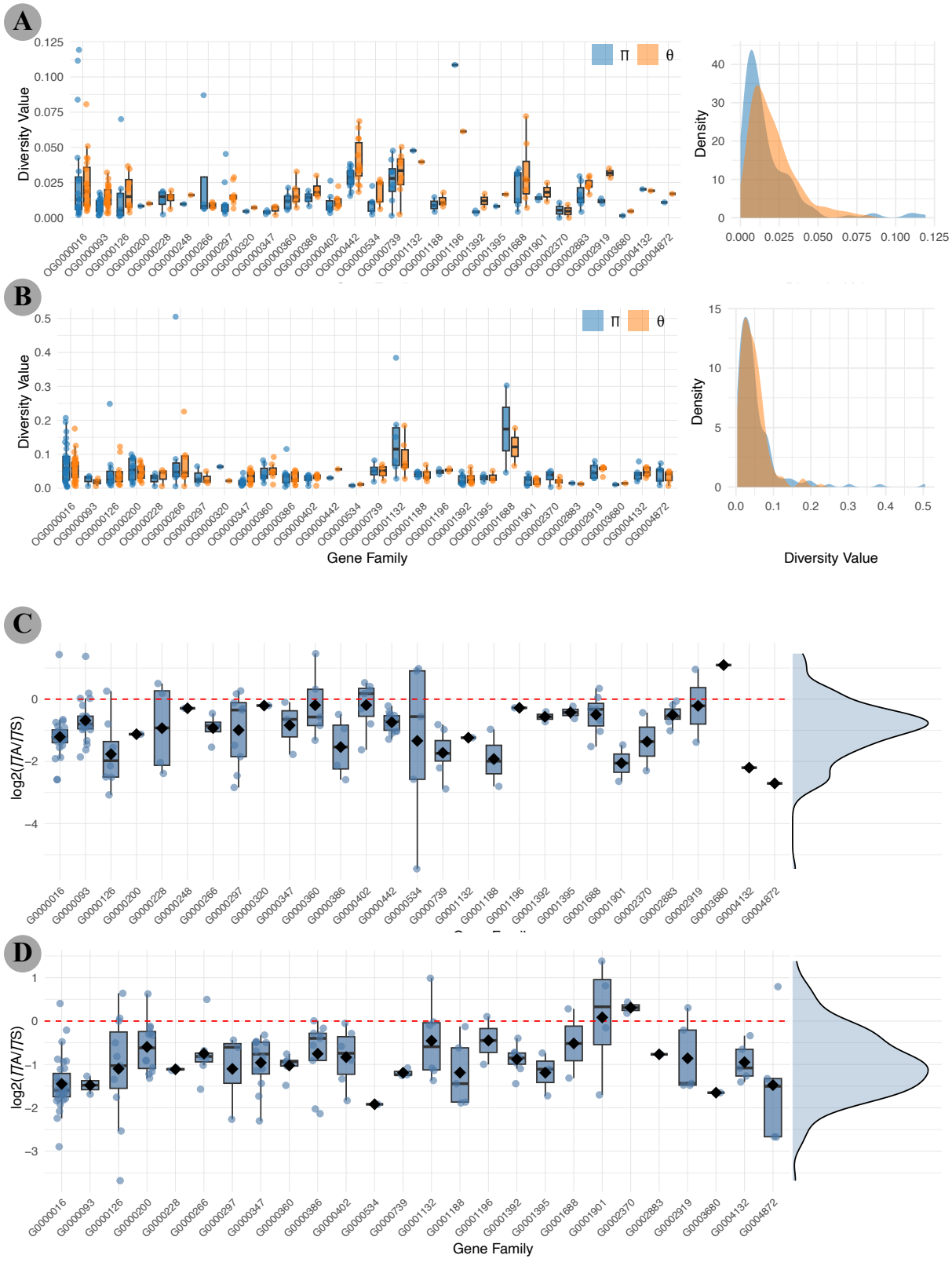

Figure 6: **Nucleotide diversity in defense- and stress-response gene families.** Boxplots show observed nucleotide diversity (π) and expected nucleotide diversity (θ_W) for *A. thaliana* **(A)** and *A. lyrata* **(B)**. Density plots to the right of each boxplot panel illustrate the overall distribution of diversity values across all gene families. Boxplots of the ratio of nonsynonymous (π_A) to synonymous (π_S) polymorphism, expressed as $\log_2(\pi_A/\pi_S)$, are shown for *A. thaliana* **(C)** and *A. lyrata* **(D)**. Positive values indicate higher nonsynonymous diversity ($\pi_A > \pi_S$); Diamonds indicate the mean values. Corresponding density plots display the distribution of ratios across all gene families, with the x-axis representing density.

do not lead to dramatic changes in gene family size over short evolutionary timescales. This steady-state model of gene family evolution has been proposed to reflect a balance between the generation of new gene copies via duplication and their subsequent loss or pseudogenization (Lynch and Conery, 2000; Hahn et al., 2007; Innan and Kondrashov, 2010).

Importantly, the robustness of our turnover rate estimates depends on the accuracy of gene family annotations. Although annotation errors such as gene model fragmentation, pseudogene misannotation, or copy number collapse may affect individual gene families, previous simulation studies suggest that turnover models are relatively robust to moderate levels of annotation error, especially when analyses are conducted across hundreds to thousands of gene families (Han et al., 2013; Hahn et al., 2005). Moreover, our analysis benefits from using high-quality genome assemblies and annotations available for the Arabidopsis species group, minimizing potential biases due to systematic errors in gene family inference.

4.3 Fast-Evolving Gene Families as Substrates of Adaptive Evolution

Although most gene families exhibit steady turnover rates, a subset of gene families shows evidence of rapid expansion and contraction, suggesting potential roles in lineage-specific adaptation. Our analysis identified several fast-evolving gene families that deviate significantly from the background birth-death process, consistent with previous observations that gene family evolution can be highly heterogeneous across the genome (Han et al., 2013; Demuth and Hahn, 2009).

Functional enrichment analyses of these rapidly evolving gene families reveal overrepresentation of categories related to defense responses, stress tolerance, regulatory functions, and multifunctional protein complexes (see Fig. 2). Specifically, gene families involved in pathogen recognition (e.g.,Thionin proteins, NB-LRR genes), abiotic stress responses (e.g., heat shock proteins, detoxification enzymes), and transcriptional regulation (e.g., transcription factor families) are among those showing elevated copy number variation. Similar enrichment patterns have been reported across multiple plant lineages, highlighting the importance of copy number dynamics in genes mediating biotic and abiotic interactions (Roulin et al., 2013; Edger and Pires, 2009).

The frequent involvement of defense- and stress-related gene families in rapid copy number evolution is consistent with the predictions of the Red Queen hypothesis and ongoing coevolutionary arms races between hosts and their pathogens (Valen, 2014; Meyers et al., 2005; Michelmore and Meyers, 1998). These gene families often experience strong selective pressures from fluctuating biotic and abiotic environments, which can favor both the expansion of new paralogs to diversify recognition capacities and the loss of obsolete or costly gene copies (Mondragón-Palomino et al., 2002; Manara et al., 2020). Copy number variation thus provides a flexible substrate for rapid evolutionary responses to shifting selective landscapes.

The observation that many of these gene family expansions are lineage-specific further supports the role of gene duplication as a mechanism for recent and ongoing adaptation. Lineage-specific expansions reflect both the differential ecological challenges faced by each species and the unique evolutionary trajectories resulting from demographic, environmental, and ecological histories (Han et al., 2013; Eric Schranz et al., 2012; Flagel and Wendel, 2009). Together, these findings reinforce the view that while most gene families evolve under steady turnover, subsets of fast-evolving gene families serve as important drivers of adaptive evolution in response to both ancient and ongoing selective pressures.

4.4 Functional Divergence, Pseudogenization, and Genome Organization

4.4.1 Functional Divergence Within Gene Families

A substantial proportion of gene families in our dataset ($> 50\%$) exhibit evidence of functional divergence among their members. This pattern is consistent with the widely accepted duplication-divergence-subfunctionalization model, in which gene duplication events are followed by sequence divergence, allowing paralogs to acquire specialized or novel functions (Ohno, 1970; Force et al., 1999; Innan and Kondrashov, 2010). Larger gene families, in particular, tend to show greater functional diversification, as the accumulation of multiple duplicates provides opportunities for both neofunctionalization and subfunctionalization (Lynch and Conery, 2000; Flagel and Wendel, 2009). This expansion of functional repertoires may contribute to species' abilities to adapt to a variety of ecological conditions and environmental stresses.

4.4.2 Pseudogenization Patterns Reflect Evolutionary Trade-offs

Our analysis further shows that a majority of expanded gene families contain pseudogenized paralogs, indicating that many duplicated genes experience relaxed selective constraints after duplication. The coexistence of functional and pseudogenized copies within the same family supports evolutionary models of dosage balance, buffering, and evolutionary experimentation (Lynch and Force, 2000; Zhang, 2003). While some duplicates are retained due to selective advantages conferred by increased gene dosage or functional redundancy, others become pseudogenes as they accumulate deleterious mutations in the absence of strong purifying selection. This dynamic balance between preservation and loss shapes long-term gene family evolution (Scannell et al., 2007; Smet et al., 2013).

4.4.3 Chromosomal Distribution Highlights Different Expansion Mechanisms

The chromosomal distribution of gene duplicates provides further insights into the mechanisms driving gene family expansions. In *A. lyrata*, gene duplicates show more extensive dispersal across the genome, often associated with transposable element (TE) insertions and genomic rearrangements. Such patterns suggest that TE-mediated transposition and ectopic recombination have played significant roles in relocating gene copies throughout the genome (Quesneville, 2020). In contrast, *A. thaliana* shows a higher proportion of tandemly duplicated gene families, consistent with localized duplication events through unequal crossing-over and replication slippage (Vision et al., 2000). These contrasting patterns may reflect differences in genome size, TE activity, and recombination landscape between the two species.

4.4.4 The Role of Transposable Elements in Shaping Gene Family Evolution

Transposable elements appear to play a major role in shaping gene family dynamics, particularly in *A. lyrata*, which exhibits both higher TE content and greater TE-association with duplicated gene copies (see Fig. 3). TE-mediated mechanisms can facilitate gene duplication, promote gene relocation, and contribute to pseudogenization through insertional mutagenesis (Feschotte and Pritham, 2007; Bennetzen and Wang, 2014). Moreover, TE-driven structural changes may create opportunities for regulatory rewiring and expression divergence among paralogs, contributing to functional diversification (Rebollo et al., 2012). The interplay between TE activity and gene duplication underscores the broader connection between genome instability and adaptive evolution, where bursts of TE activity may occasionally generate evolutionary innovations that become substrates for natural selection (Oliver and Greene, 2009).

4.5 Copy Number Variation Is Geographically Structured

Our analysis demonstrates that copy number variation (CNV) exhibits clear patterns of population-level and geographic structuring across both *Arabidopsis thaliana* and *A. lyrata*. Multivariate analyses of CNV profiles, including multiple correspondence analysis (MCA), reveal well-defined geographic clusters (see Fig. 4), indicating that CNV differentiation is not randomly distributed across populations but instead reflects underlying biogeographic and demographic processes.

In *A. thaliana*, CNV variation aligns closely with previously described population lineages, particularly the distinction between relict populations and the widespread non-relict (post-glacial expansion) groups (François et al., 2008; Mitchell-Olds and Schmitt, 2006). Relict populations, often found in southern European refugia, retain ancient genomic variation that includes distinct CNV profiles, while non-relict populations show evidence of CNV patterns shaped by more recent range expansions and associated bottlenecks. These results suggest that demographic history, including glacial refugia and post-glacial colonization, has contributed substantially to shaping CNV diversity within *A. thaliana* (Alonso-Blanco et al., 2016). In addition to this, we also observe that peripheral populations, such as those from the Cape Verde Islands and Madeira, exhibit unique CNV profiles that may reflect local adaptation to distinct environmental conditions or isolation from continental gene flow (Wlodzimierz et al., 2023; Lian et al., 2024).

Similarly, in *A. lyrata*, geographic structuring of CNV reflects broader continental divergence patterns observed across its circumboreal distribution (Scott et al., 2024; Hämälä et al., 2018b). Populations from Europe and North America exhibit distinct CNV profiles, consistent with historical divergence, limited gene flow, and adaptation to contrasting environmental conditions. The presence of geographically structured CNV in both species supports the view that copy number variation is not governed solely by stochastic mutation and genetic drift but is also influenced by population history, lineage divergence, and potentially by geographically heterogeneous selection pressures (Alkan et al., 2011; Mérot et al., 2020).

4.6 Ecological Associations of CNV Suggest Local Adaptation

Our results reveal significant and robust correlations between copy number variation (CNV) and climatic variables in both *Arabidopsis thaliana* and *A. lyrata*, consistent with the hypothesis that CNV contributes to local adaptation. In *A. thaliana*, we found that multiple temperature-related variables, particularly Bio3 (Isothermality), Bio4 (Temperature Seasonality), Bio8 (Mean Temperature of the Wettest Quarter), and Bio10 (Mean Temperature of the Warmest Quarter), exhibited strong associations with CNV. In addition, precipitation-related variables such as Bio13 (Precipitation of the Wettest Month) and Bio14 (Precipitation of the Driest Month) were also significantly associated (see Fig. 5). These associations remained statistically significant even after correcting for population structure, indicating that the observed CNV-environment correlations are not solely the result of demographic history but likely reflect adaptive responses to environmental heterogeneity.

In *A. lyrata*, CNV patterns showed significant associations with Bio2 (Mean Diurnal Range), Bio3 (Isothermality), and Bio7 (Temperature Annual Range), with all associations persisting after accounting for population structure. Unlike *A. thaliana*, the associations in *A. lyrata* were more tightly focused on temperature-related variables, suggesting differing ecological pressures or niche constraints (see Fig. 5). Notably, Siberian and North American populations formed distinct separate clusters in multivariate space, further supporting the idea that CNV variation in this species reflects adaptation to geographically structured climatic regimes.

The differences in climatic correlates between species likely stem from a combination of ecological, historical, and life-history factors, including mating system, range size, and demographic connectivity (Hämälä et al., 2018b; Exposito-Alonso et al., 2019). The influence of precipitation, especially in *A. thaliana* is particularly compelling, as precipitation seasonality

and extremes are known to shape both abiotic stress responses (e.g., drought tolerance) and biotic interactions such as pathogen pressure (Gupta et al., 2016; Szczepaniec and Finke, 2019). The persistent association between CNV and these climatic variables highlights a likely role for CNV in modulating ecological responses to both water availability and disease, particularly in environmentally heterogeneous landscapes.

4.7 A Core Set of Defense-Related Gene Families Drive Ecological Divergence

Multivariate analyses of CNV patterns, particularly MCA loadings, identified a consistent set of gene families that contribute disproportionately to ecological divergence in both *Arabidopsis thaliana* and *A. lyrata*. Despite differences in their demographic histories, both species share a core group of ecologically responsive gene families that exhibited the highest squared distances in ordination space. These include OG0000442 (plant thionin family protein), OG0000093 (cysteine-rich receptor-like kinase), and OG0000266 (WRKY transcription factor, Group II-e) in *A. thaliana*, and OG0000016 (Tetratricopeptide repeat-like protein), OG0001132 (NBS-LRR disease resistance protein), and OG0000347 (plastocyanin-like electron carrier) in *A. lyrata*.

These gene families are functionally unified by their roles in defense signaling, pathogen recognition, abiotic stress response, and regulatory flexibility-traits that are critical for plant survival under fluctuating climatic and ecological conditions. Their strong contribution to ecological divergence and consistent presence among top CNV contributors underscore the importance of defense-related genes as hotspots of genomic plasticity under selection.

Moreover, the recurrence of similar gene families across both species points toward convergent or parallel evolutionary pressures acting on CNV loci, particularly those involved in defense and stress responses (Meyers et al., 2005; Exposito-Alonso et al., 2019; Young et al., 2016). This pattern reinforces the notion that gene families linked to environmental responsiveness, especially those modulating pathogen defense and abiotic stress signaling are dynamic substrates of adaptive evolution (Eric Schranz et al., 2012; Roulin et al., 2013). Gene copy number changes in these families may offer a flexible genomic mechanism through which populations can rapidly respond to spatial and temporal variability in environmental pressures.

4.8 Selection Patterns Highlight the Evolutionary Role of Gene Duplication

Patterns of molecular evolution in gene families with copy number variation provide further insight into the evolutionary consequences of gene duplication. Multi-copy gene families exhibit elevated nucleotide diversity and higher Tajima's D values compared to single-copy genes, consistent with relaxed purifying selection acting on many duplicated loci (see Fig. 6). These patterns reflect the fact that duplicated gene copies often experience reduced selective constraint following duplication, allowing them to accumulate mutations at a faster rate relative to their single-copy counterparts (Lynch and Conery, 2000; Hahn, 2009; Smet et al., 2013). This relaxed constraint is particularly pronounced in large gene families, where some paralogs may become functionally redundant or evolve new functions, leading to increased sequence variability (Flagel and Wendel, 2009). The elevated nucleotide diversity and positive Tajima's D values observed in many duplicated gene families suggest that while most paralogs are subject to purifying selection, a subset may be evolving under relaxed constraints or even adaptive divergence. This is particularly evident in defense-related gene families, where some paralogs show signatures of positive selection, indicating that they may be involved in adaptive responses to environmental pressures such as pathogen attack or abiotic stress (del Pozo and Ramirez-Parra, 2015; Schrider and Kern, 2017). The presence of both purifying selection on essential genes and adaptive divergence in specific paralogs highlights the dual role of gene duplication in providing both robustness to core biological functions and opportunities for functional innovation (Flagel and Wendel, 2009; Kondrashov, 2012). Gene duplication thus serves as a key mechanism for generat-

ing genomic diversity and functional innovation, allowing organisms to explore new biochemical or regulatory roles while maintaining essential functions. This dual role of gene duplication is particularly evident in the context of plant defense and stress responses, where duplicated genes can provide both redundancy and flexibility in adapting to changing environmental conditions (Ohno, 1970; Zhang, 2003).

However, while many duplicated genes evolve under relaxed constraint, a subset of paralogs exhibits signatures of adaptive divergence. In these cases, positive Tajima's D values and elevated sequence polymorphism suggest that certain gene copies may be subject to balancing or directional selection, potentially reflecting their roles in environmental adaptation (Schrider and Kern, 2017; Kondrashov, 2012). This contrast between pervasive purifying selection on essential genes and adaptive divergence in specific paralogs highlights how gene duplication creates opportunities for functional experimentation while preserving core biological functions.

Gene duplication thus acts as both a source of genomic robustness and evolutionary innovation: while some copies retain ancestral functions, others are freed to explore novel biochemical or regulatory roles that may confer selective advantages under changing environmental conditions. Our results underscore the dual role of gene duplication as a fundamental process shaping both the structural diversity of plant genomes and their capacity for ecological and adaptive diversification.

5 Conclusion

In summary, our study reveals that while most gene families in Arabidopsis evolve under steady turnover, a subset, particularly those involved in defense and stress responses exhibits rapid, lineage-specific CNV dynamics shaped by geographic structure and environmental gradients. These CNV patterns are associated with local adaptation, persist after controlling for population structure, and are enriched in gene families showing functional divergence and signatures of relaxed or adaptive selection. Together, our findings highlight gene duplication and CNV as key mechanisms driving ecological divergence and evolutionary flexibility in plant genomes.

Acknowledgment

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 456082119 - TRR 341, in Cologne and Düsseldorf.

Data Sharing

The reference genomes used in the analysis are from *A. thaliana*: TAIR10: GCA_000001735.1 Cheng et al. (2017); *A. lyrata*: GCA_000004255.1Hu et al. (2011); Rawat et al. (2015); *A. halleri*: Arabidopsis_halleri v2.03, DOE-JGI ; *A. alpina*: GCA_000733195.1; and *Carica papaya*: Carica_papaya ASGPBv0.4 Ming et al. (2008); with their corresponding annotations. The long-read sequences of *A. thaliana* were downloaded from PRJNA777107 , PRJNA779205 PRJNA834751, PRJNA1033522, PRJEB55353, PRJEB50694, and PRJEB55632 NCBI bioprojects with their corresponding meta-data. Genome assemblies of *A. thaliana* are also from above NCBI bioprojects. The *A. lyrata* long-read genome assemblies of *A. lyrata* are deposited at `TBA` and can be accessed on request. The *A. lyrata* SNP dataset of Illumina WGS are from Scott et al. (2024) and can be accessed at `TBA`. All the scripts and intermediate data used in the analysis are archived in a GitHub repository and can be accessed at `https://github.com/piyalkarum/GeneFamilyEvolution`.

References

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–376.

Almaghrabi, B., Ali, M. A., Zahoor, A., Shah, K. H., and Bohlmann, H. (2019). Arabidopsis thionin-like genes are involved in resistance against the beet-cyst nematode (Heterodera schachtii). *Plant Physiology and Biochemistry*, 140:55–67.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., Horton, M., Jarsulic, M., Kerstetter, R. A., Korte, A., Korte, P., Lanz, C., Lee, C.-R., Meng, D., Michael, T. P., Mott, R., Muliyati, N. W., Nägele, T., Nagler, M., Nizhynska, V., Nordborg, M., Novikova, P. Y., Picó, F. X., Platzer, A., Rabanal, F. A., Rodriguez, A., Rowan, B. A., Salomé, P. A., Schmid, K. J., Schmitz, R. J., Ümit Seren, Sperone, F. G., Sudkamp, M., Svardal, H., Tanzer, M. M., Todd, D., Volchenboum, S. L., Wang, C., Wang, G., Wang, X., Weckwerth, W., Weigel, D., and Zhou, X. (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, 166(2):481–491.

Bennetzen, J. L. and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual Review of Plant Biology*, 65(Volume 65, 2014):505–530.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2019). The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis*, 57(1):e23295.

Brummitt, R. K., Pando, F., Hollis, S., and Brummitt, N. A. (2001). *World Geographical Scheme for Recording Plant Distributions*. Hunt Institute for Botanical Documentation, Carnegie Mellon University, Pittsburgh, PA, 2 edition.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009a). Blast+: Architecture and applications. *BMC Bioinformatics*, 10:421.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009b). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421.

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *arabidopsis thaliana*. *BMC Plant Biology*, 4(1):10.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., and Weigel, D. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, 43(10):956–963.

Cheng, C.-Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., and Town, C. D. (2017). Araport11: a complete reannotation of the arabidopsis thaliana reference genome. *The Plant Journal*, 89(4):789–804.

Consortium, T. U. (2022). Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531.

del Pozo, J. C. and Ramirez-Parra, E. (2015). Whole genome duplications in plants: an overview from *Arabidopsis*. *Journal of Experimental Botany*, 66(22):6991–7003.

Demuth, J. P., Bie, T. D., Stajich, J. E., Cristianini, N., and Hahn, M. W. (2006). The evolution of mammalian gene families. *PLOS ONE*, 1(1):1–10.

Demuth, J. P. and Hahn, M. W. (2009). The life and death of gene families. *BioEssays*, 31(1):29–39.

Edger, P. P. and Pires, J. C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research*, 17(5):699–717.

Emms, D. M. and Kelly, S. (2015). Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16:157.

Emms, D. M. and Kelly, S. (2019). Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1):238.

Eric Schranz, M., Mohammadin, S., and Edger, P. P. (2012). Ancient whole genome duplications, novelty and diversification: the wgd radiation lag-time model. *Current Opinion in Plant Biology*, 15(2):147–153. Genome studies molecular genetics.

Exposito-Alonso, M., Gómez Rodríguez, R., Barragán, C., Capovilla, G., Chae, E., Devos, J., Dogan, E. S., Friedemann, C., Gross, C., Lang, P., Lundberg, D., Middendorf, V., Kageyama, J., Karasov, T., Kersten, S., Petersen, S., Rabbani, L., Regalado, J., Reinelt, L., Rowan, B., Seymour, D. K., Symeonidi, E., Schwab, R., Tran, D. T. N., Venkataramani, K., Van de Weyer, A.-L., Vasseur, F., Wang, G., Wedegärtner, R., Weiss, F., Wu, R., Xi, W., Zaidem, M., Zhu, W., García-Arenal, F., Burbano, H. A., Bossdorf, O., Weigel, D., Nielsen, R., and 500 Genomes Field Experiment Team (2019). Natural selection on the *Arabidopsis thaliana* genome in present and future climates. *Nature*, 573(7772):126–129.

Feschotte, C. and Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics*, 41(Volume 41, 2007):331–368.

Flagel, L. E. and Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytologist*, 183(3):557–564.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545.

François, O., Blum, M. G. B., Jakobsson, M., and Rosenberg, N. A. (2008). Demographic history of european populations of *Arabidopsis thaliana*. *PLOS Genetics*, 4(5):1–15.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., Carter, N. P., Scherer, S. W., and Lee, C. (2006). Copy number variation: new insights in genome diversity. *Genome Research*, 16(8):949–961. Epub 2006 Jun 29.

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, 201(4):1555–1579.

Gupta, A., Sarkar, A. K., and Senthil-Kumar, M. (2016). Global transcriptional analysis reveals unique and shared responses in *Arabidopsis thaliana* exposed to combined drought and pathogen stress. *Frontiers in Plant Science*, Volume 7 - 2016.

Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity*, 100(5):605–617.

Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C., and Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, 15(8):1153–1160.

Hahn, M. W., Han, M. V., and Han, S.-G. (2007). Gene family evolution across 12 *Drosophila* genomes. *PLOS Genetics*, 3(11):1–12.

Han, M. V., Thomas, G. W., Lugo-Martinez, J., and Hahn, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution*, 30(8):1987–1997.

Hanikenne, M., Talke, I. N., Haydon, M. J., Lanz, C., Nolte, A., Motte, P., Kroymann, J., Weigel, D., and Krämer, U. (2008). Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. *Nature*, 453(7193):391–395.

Hendriks, K. P., Kiefer, C., Al-Shehbaz, I. A., Bailey, C. D., Hooft van Huysduynen, A., Nikolov, L. A., Nauheimer, L., Zuntini, A. R., German, D. A., Franzke, A., Koch, M. A., Lysak, M. A., Óscar Toro-Núñez, Özüdoğru, B., Invernón, V. R., Walden, N., Maurin, O., Hay, N. M., Shushkov, P., Mandáková, T., Schranz, M. E., Thulin, M., Windham, M. D., Rešetnik, I., Španiel, S., Ly, E., Pires, J. C., Harkess, A., Neuffer, B., Vogt, R., Bräuchler, C., Rainer, H., Janssens, S. B., Schmull, M., Forrest, A., Guggisberg, A., Zmarzty, S., Lepschi, B. J., Scarlett, N., Stauffer, F. W., Schönberger, I., Heenan, P., Baker, W. J., Forest, F., Mummenhoff, K., and Lens, F. (2023). Global Brassicaceae phylogeny based on filtering of 1,000-gene dataset. *Current Biology*, 33(19):4052–4068.e6.

Hohmann, N., Wolf, E. M., Lysak, M. A., and Koch, M. A. (2015). A time-calibrated road map of brassicaceae species radiation and evolutionary history. *The Plant Cell*, 27(10):2770–2784.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.-F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F. X., Van de Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., and Guo, Y.-L. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics*, 43(5):476–481.

Hämälä, T., Mattila, T. M., Leinonen, P. H., Kuittinen, H., and Savolainen, O. (2018a). Role of seed germination in adaptation and reproductive isolation in *Arabidopsis lyrata*. *Molecular Ecology*, 27(3):1004–1017.

Hämälä, T., Mattila, T. M., and Savolainen, O. (2018b). Local adaptation and ecological differentiation under selection, migration, and drift in *Arabidopsis lyrata**. *Evolution*, 72(7):1373–1386.

Innan, H. and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97–108.

Kang, Y. J., Kim, K. H., Shim, S., Yoon, M. Y., Sun, S., Kim, M. Y., Van, K., and Lee, S.-H. (2012). Genome-wide mapping of NBS-LRR genes and their association with disease resistance in soybean. *BMC Plant Biology*, 12(1):139.

Kapitonov, V. V. and Jurka, J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends in Genetics*, 23(10):521–529. PMID: 17850916.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., and Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, 4(1):170122.

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences*, 279(1749):5048–5057.

Lian, Q., Huettel, B., Walkemeier, B., Mayjonade, B., Lopez-Roques, C., Gil, L., Roux, F., Schneeberger, K., and Mercier, R. (2024). A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range. *Nature Genetics*, 56(5):982–991.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155.

Lynch, M. and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473.

Ma, S., Niu, J., Si, Y., Zheng, S., Lu, Y., Tian, S., Shi, X., Chen, Z., Sun, C., Qin, Z., Liu, X., Wu, H., Gu, M., Cui, M., Lu, Q., Zhou, W., He, W., Zhang, C., He, F., and Ling, H.-Q. (2025). A comprehensive map of DNA-segment copy number variation in 491 genomes of common wheat uncovers genes associated with multiple agronomic traits. *Plant Communications*, 6(3):101226.

Manara, A., Fasani, E., Furini, A., and DalCorso, G. (2020). Evolution of the metal hyperaccumulation and hypertolerance traits. *Plant, Cell & Environment*, 43(12):2969–2986.

Maron, L. G., Guimarães, C. T., Kirst, M., Albert, P. S., Birchler, J. A., Bradbury, P. J., Buckler, E. S., Coluccio, A. E., Danilova, T. V., Kudrna, D., Magalhaes, J. V., Piñeros, M. A., Schatz, M. C., Wing, R. A., and Kochian, L. V. (2013). Aluminum tolerance in maize is associated with higher *MATE1* gene copy number. *Proceedings of the National Academy of Sciences*, 110(13):5241–5246.

Martin, F. J., Amode, M. R., Aneja, A., Austine-Orimoloye, O., Azov, A., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S., Bignell, A., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Charkhchi, M., Cockburn, A., Da Rin Fiorretto, L., Davidson, C., Dodiya, K., Donaldson, S., El Houdaigui, B., El Naboulsi, T., Fatima, R., Giron, C. G., Genez, T., Ghattaoraya, G. S., Martinez, J. G., Guijarro, C., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Kay, M., Kaykala, V., Le, T., Lemos, D., Marques-Coelho, D., Marugán, J. C., Merino, G., Mirabueno, L., Mushtaq, A., Hossain, S., Ogeh, D. N., Sakthivel, M. P., Parker, A., Perry, M., Piližota, I., Prosovetskaia, I., Pérez-Silva, J. G., Salam, A., Saraiva-Agostinho, N., Schuilenburg, H., Sheppard, D., Sinha, S., Sipos, B., Stark, W., Steed, E., Sukumaran, R., Sumathipala, D., Suner, M.-M., Surapaneni, L., Sutinen, K., Szpak, M., Tricomi, F., Urbina-Gómez, D., Veidenberg, A., Walsh, T., Walts, B., Wass, E., Willhoft, N., Allen, J., Alvarez-Jarreta, J., Chakiachvili, M., Flint, B., Giorgetti, S., Haggerty, L., Ilsley, G., Loveland, J., Moore, B., Mudge, J., Tate, J., Thybert, D., Trevanion, S., Winterbottom, A., Frankish, A., Hunt, S. E., Ruffier, M., Cunningham, F., Dyer, S., Finn, R., Howe, K., Harrison, P. W., Yates, A. D., and Flicek, P. (2022). Ensembl 2023. *Nucleic Acids Research*, 51(D1):D933–D941.

Mendes, F. K., Vanderpool, D., Fulton, B., and Hahn, M. W. (2020). Cafe 5 models variation in evolutionary rates among gene families. *Bioinformatics*, 36(22-23):5516–5518.

Meyers, B. C., Kaushik, S., and Nandety, R. S. (2005). Evolving disease resistance genes. *Current Opinion in Plant Biology*, 8(2):129–134.

Michelmore, R. W. and Meyers, B. C. (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Research*, 8(11):1113–1130.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., Senin, P., Wang, W., Ly, B. V., Lewis, K. L. T., Salzberg, S. L., Feng, L., Jones, M. R., Skelton, R. L., Murray, J. E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R. E., Michael, T. P., Wall, K., Rice, D. W., Albert, H., Wang, M.-L., Zhu, Y. J., Schatz, M., Nagarajan, N., Acob, R. A., Guan, P., Blas, A., Wai, C. M., Ackerman, C. M., Ren, Y., Liu, C., Wang, J., Wang, J., Na, J.-K., Shakirov, E. V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X., Bowers, J. E., Gschwend, A. R., Delcher, A. L., Singh, R., Suzuki, J. Y., Tripathi, S., Neupane, K., Wei, H., Irikura, B., Paidi, M., Jiang, N., Zhang, W., Presting, G., Windsor, A., Navajas-Pérez, R., Torres, M. J., Feltus, F. A., Porter, B., Li, Y., Burroughs, A. M., Luo, M.-C., Liu, L., Christopher, D. A., Mount, S. M., Moore, P. H., Sugimura, T., Jiang, J., Schuler, M. A., Friedman, V., Mitchell-Olds, T., Shippen, D. E., dePamphilis, C. W., Palmer, J. D., Freeling, M., Paterson, A. H., Gonsalves, D., Wang, L., and Alam, M. (2008). The draft genome of the transgenic tropical fruit tree papaya (carica papaya linnaeus). *Nature*, 452(7190):991–996.

Mitchell-Olds, T. and Schmitt, J. (2006). Genetic mechanisms and evolutionary significance of natural variation in *Arabidopsis*. *Nature*, 441(7096):947–952.

Mondragón-Palomino, M., Meyers, B. C., Michelmore, R. W., and Gaut, B. S. (2002). Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Research*, 12(9):1305–1315.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., and Rafalski, A. (2005). Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics*, 37(9):997–1002.

Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H., Crollius, H. R., and Salse, J. (2015). Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biology*, 16(1):262.

Muñoz-Amatriaín, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., Scholz, U., Ariyadasa, R., Spannagl, M., Nussbaumer, T., Mayer, K. F., Taudien, S., Platzer, M., Jeddeloh, J. A., Springer, N. M., Muehlbauer, G. J., and Stein, N. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biology*, 14(6):R58.

Mérot, C., Oomen, R. A., Tigano, A., and Wellenreuther, M. (2020). A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*, 35(7):561–572.

Nei, M. and Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics*, 39(Volume 39, 2005):121–152.

Novikova, P. Y., Hohmann, N., Nizhynska, V., Tsuchimatsu, T., Ali, J., Muir, G., Guggisberg, A., Paape, T., Schmid, K., Fedorenko, O. M., Holm, S., Säll, T., Schlötterer, C., Marhold, K., Widmer, A., Sese, J., Shimizu, K. K., Weigel, D., Krämer, U., Koch, M. A., and Nordborg, M. (2016). Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics*, 48(9):1077–1082.

Ohno, S. (1970). *Evolution by Gene Duplication*. Springer-Verlag, Berlin, Heidelberg.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2022). *vegan: Community Ecology Package*. R package version 2.6-4.

Oliver, K. R. and Greene, W. K. (2009). Transposable elements: powerful facilitators of evolution. *BioEssays*, 31(7):703–714.

Otto, M. and Wiehe, T. (2023). The structured coalescent in the context of gene copy number variation. *Theoretical Population Biology*, 154:67–78.

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. *Plant Physiology*, 171(4):2294–2316.

Paradis, E. and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528.

Prunier, J., Caron, S., Lamothe, M., Blais, S., Bousquet, J., Isabel, N., and MacKay, J. (2017). Gene copy number variations in adaptive evolution: The genomic distribution of gene copy number variations revealed by genetic mapping and their adaptive role in an undomesticated species, white spruce (*Picea glauca*). *Molecular Ecology*, 26(21):5989–6001.

Quesneville, H. (2020). Twenty years of transposable element analysis in the *arabidopsis thaliana* genome. *Mobile DNA*, 11(1):28.

Rawat, V., Abdelsamad, A., Pietzenuk, B., Seymour, D. K., Koenig, D., Weigel, D., Pecinka, A., and Schneeberger, K. (2015). Improving the annotation of arabidopsis lyrata using rna-seq data. *PLOS ONE*, 10(9):1–12.

Rebollo, R., Romanish, M. T., and Mager, D. L. (2012). Transposable elements: An abundant and natural source of regulatory sequences for host genes. *Annual Review of Genetics*, 46(Volume 46, 2012):21–42.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–454.

Rizzon, C., Ponger, L., and Gaut, B. S. (2006). Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. *PLOS Computational Biology*, 2(9):1–12.

Ross-Ibarra, J., Wright, S. I., Foxe, J. P., Kawabe, A., DeRose-Wilson, L., Gos, G., Charlesworth, D., and Gaut, B. S. (2008). Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLOS ONE*, 3(6):1–13.

Roulin, A., Auer, P. L., Libault, M., Schlueter, J., Farmer, A., May, G., Stacey, G., Doerge, R. W., and Jackson, S. A. (2013). The fate of duplicated genes in a polyploid plant genome. *The Plant Journal*, 73(1):143–153.

Scannell, D. R., Frank, A. C., Conant, G. C., Byrne, K. P., Woolfit, M., and Wolfe, K. H. (2007). Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proceedings of the National Academy of Sciences*, 104(20):8397–8402.

Schrider, D. R. and Hahn, M. W. (2010). Gene copy-number polymorphism in nature. *Proceedings of the Royal Society B: Biological Sciences*, 277(1698):3213–3221.

Schrider, D. R. and Kern, A. D. (2017). Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular Biology and Evolution*, 34(8):1863–1877.

Scott, A. D., Kolesnikova, U., Glushkevich, A., Steinmann, L., Tikhomirov, N., Pfordt, U., Bohutínská, M., Burns, R., Seregin, A. P., Kolar, F., Schmickl, R., and Novikova, P. Y. (2024). Multiple polyploidizations in *Arabidopsis lyrata* stabilized by long-range adaptive introgression across Eurasia. *bioRxiv*.

Smet, R. D., Adams, K. L., Vandepoele, K., Montagu, M. C. E. V., Maere, S., and de Peer, Y. V. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, 110(8):2898–2903.

Stiehler, F., Steinborn, M., Scholz, S., Dey, D., Weber, A. P. M., and Denton, A. K. (2020). Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics*, 36(22-23):5291–5298.

Szczepaniec, A. and Finke, D. (2019). Plant-vector-pathogen interactions in the context of drought stress. *Frontiers in Ecology and Evolution*, Volume 7 - 2019.

Todesco, M., Balasubramanian, S., Hu, T. T., Traw, M. B., Horton, M., Epple, P., Kuhns, C., Sureshkumar, S., Schwartz, C., Lanz, C., Laitinen, R. A. E., Huang, Y., Chory, J., Lipka, V., Borevitz, J. O., Dangl, J. L., and Weigel, D. (2010). Natural allelic variation underlying a major fitness trade-off in *Arabidopsis thaliana*. *Nature*, 465:632–636.

Valen, L. V. (2014). A new evolutionary law (1973). In Smith, F. A., Gittleman, J. L., and Brown, J. H., editors, *Foundations of Macroecology*, pages 284–314. University of Chicago Press, Chicago.

Vision, T. J., Brown, D. G., and Tanksley, S. D. (2000). The origins of genomic duplications in *Arabidopsis*. *Science*, 290(5499):2114–2117.

Wang, Q., Wang, M., Zhang, X., Hao, B., Kaushik, S. K., and Pan, Y. (2011). Wrky gene family evolution in arabidopsis thaliana. *Genetica*, 139(8):973–983.

Wang, Y., Wang, X., and Paterson, A. H. (2012). Genome and gene duplications and gene expression divergence: a view from plants. *Annals of the New York Academy of Sciences*, 1256(1):1–14.

Wlodzimierz, P., Rabanal, F. A., Burns, R., Naish, M., Primetis, E., Scott, A., Mandáková, T., Gorringe, N., Tock, A. J., Holland, D., Fritschi, K., Habring, A., Lanz, C., Patel, C., Schlegel, T., Collenberg, M., Mielke, M., Nordborg, M., Roux, F., Shirsekar, G., Alonso-Blanco, C., Lysak, M. A., Novikova, P. Y., Bousios, A., Weigel, D., and Henderson, I. R. (2023). Cycles of satellite and transposon evolution in *Arabidopsis* centromeres. *Nature*, 618(7965):557–565.

Young, N. D., Zhou, P., and Silverstein, K. A. (2016). Exploring structural variants in environmentally sensitive gene families. *Current Opinion in Plant Biology*, 30:19–24.

Zhang, F., Carvalho, C. M., and Lupski, J. R. (2009). Complex human chromosomal and genomic rearrangements. *Trends in Genetics*, 25(7):298–307.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6):292–298.