DupyliCate: mining, classifying, and characterizing gene duplications

Shakunthala Natarajan[1] & Boas Pucker[1*]

[1] Plant Biotechnology and Bioinformatics, Institute for Cellular and Molecular Botany, University of Bonn, Kirschallee 1, 53115, Bonn, Germany

* Corresponding author: pucker@uni-bonn.de

Abstract

Paralogs, copies of a gene, form an important basis for novelty during evolution. Analysis of such gene duplications is important to understand the emergence of novel traits during evolution. DupyliCate is a Python tool that has been developed for this purpose. With the ability to process multiple datasets concurrently, flexible features and parameters to set species-specific thresholds, DupyliCate offers a high throughput method for gene copy identification and analysis. The different available parameters and modes are explored in detail based on the *Arabidopsis thaliana* datasets. Proof of concept for the tool is presented by characterizing well known duplications in different plants, and its broad applicability is demonstrated by running it on diverse datasets including complex plant genome sequences with high heterozygosity. Further, two case studies involving the evolution of *Flavonol synthase* (*FLS*) genes in Brassicales, and the evolution of flavonol synthesis regulating myeloblastosis (MYB) genes - *MYB12* and *MYB111* across a large number of plant species, are presented as exemplar use cases. The tool's applicability beyond plants is demonstrated on *Escherichia coli, Saccharomyces cerevisiae, and Caenorhabditis elegans* datasets. DupyliCate is available at: https://github.com/ShakNat/DupyliCate.

Key words: Gene duplications, comparative genomics, expression analysis, high throughput, Ka/Ks, evolution

Introduction

Gene duplication products also called paralogs are homologous genes originating from a common ancestor gene. In contrast to orthologs, they arise through duplication events and not speciation events[1]. They are an important source of evolutionary novelty in all organisms and are major drivers for the diversification of biosynthetic pathways seen in plants and other organisms. They can broadly be divided into small-scale duplications and large-scale duplications. The small-scale duplication group encompasses tandem duplications, proximal

duplications, transposon-derived duplications, and retroduplications. Whereas, the large scale duplication group includes whole genome duplications, and segmental duplications[2]. Identification of gene duplications is fraught with difficulties like functional redundancy of gene duplicates, and extensive small scale duplication events. It is also compounded by gene loss, rearrangement, mutation events across evolutionary time scales and presence of multidomain protein structures[2,3]. Lastly, gene duplicates can be formed by a number of mechanisms, and have different ages and evolutionary fates, further increasing the difficulty of detection and analysis[1].

Nevertheless, there have been huge research efforts in this direction resulting in a plethora of tools and methodologies to identify gene duplications. GenomeHistory[4], was among the first tools to provide a basic framework for gene duplication detection with a combination of local and global alignment approaches followed by pairwise computation of Ka/Ks values to evaluate selection pressures. But this software's web page no longer exists and is hence not accessible. Following this, several tools like OrthoMCL[5], OrthoFinder[6], MCScanX[7] were developed to identify orthologs and as a result also distinguish them from paralogs using tree-based approaches and synteny approaches, respectively. All these tools focus more on the ortholog identification aspects. Next, DuplicationDetector[8] was a Perl script developed to identify putative recent segmental duplicates in homozygous organisms from next generation sequencing data. It focuses on identifying large scale duplications like segmental duplications, at the DNA level. Later, with a detailed study on gene duplications and evolution in plants, DupGen_finder[9], a Perl based script was developed to identify gene duplicate pairs, and comprehensively classify gene duplicates into tandem duplicates, proximal duplicates, transposed duplicates and whole genome duplicates. The study also provided separate scripts for Ka/Ks analysis by integrating KaKs_Calculator 2.0[10], a dedicated tool for Ka/Ks calculations. The work further involved expression divergence analysis of the identified gene duplicates, but this analysis was not a part of the DupGen_finder main script, and is not expandable to wide-ranging user data. Zhang and co-workers later developed HSDFinder[11], a BLAST-based web tool to annotate, categorize and visualize highly similar duplicates across species. More recently, an updated version of this tool, called HSDSnake[12] was released as a user-friendly SnakeMake pipeline, with added features to classify duplicate pairs by integrating the DupGen_finder script and compute Ka/Ks ratios. WGDI[13] was another Python-based command line tool that was developed to identify and characterize whole genome duplication related events in species, thus focusing on large scale duplications. Another Python-based tool integrating machine learning framework to detect orthologs, annotate them, identify gene losses and duplications across alignable genomes is TOGA (Tool to infer Orthologs from Genome Alignments)[14]. TOGA leverages alignments between intergenic regions and introns to identify orthologs, and perform other auxiliary analyses. Following this, doubletrouble[15] was released as an R package to classify and identify gene duplications. It works on a framework similar to DupGen_finder, identifies duplicate pairs, classifies them, calculates Ka/Ks ratios, but also goes one step ahead in helping identify retrotransposon derived duplicates and DNA transposon derived duplicates. Finally, SegmantX[16] is another recent Python script that uses a novel approach of local alignment

followed by a chaining algorithm to identify segmental duplications at the DNA level and also provides a docker container for ease of use.

From the varying tools discussed above, the following aspects are evident - first, most of the gene duplication identification tools identify duplicates in pairs. But in a biological context there can be arrays of duplicates and identifying them all and maintaining the link between them is crucial for further evolutionary understanding. Next, most tools have been adapted for use on standardized data from databases like Ensembl and NCBI, limiting the use cases on widely varying bioinformatics data formats. Given the heterogeneity of General Feature Format (GFF) files from different sources, and databases, it is necessary to come up with tools that can work across different GFF file formats. Furthermore, many tools focus on either identifying large scale duplications and characterizing relevant evolutionary events or classifying duplications. But it is also important to integrate expression analysis, to understand the functional divergence of identified duplicates and also study the nature of the identified duplicates by adding details on genomic order conservation with synteny, in cases of comparative genomic studies between two species. Finally, from a biological perspective, different species have different duplication landscapes in their genomes. However, existing methods work based on defined thresholds necessitating the use of species-specific thresholds[2].

In this context, DupyliCate is presented here as a solution for gene duplication array identification, and broad classification of the duplicates. It provides options for downstream expression analysis of the identified duplicates to understand the expression divergence, and gene-level Ka/Ks computations implemented independently as a part of the script, to infer selection pressures. It has been deployed here on standard datasets from NCBI, phytozome, and varying additional sources, to demonstrate its ability to process widely varying input file formats. It also leverages BUSCO-based metrics to determine species-specific thresholds for gene duplicates identification, and offers the users a wide range of parameters that can be fine-tuned according to the objectives at hand.

Results

Benchmarking results

DupyliCate was compared against the tools doubletrouble and DupGen_finder. The benchmarking results of DupyliCate runs using manual threshold and BUSCO-based threshold for singleton-duplicate segregation were different from one another as well from the other tools compared (Table 1). The run with manual thresholds (a) gave numbers comparable to the results output by both doubletrouble and DupGen_finder. The number of gene duplicates classified by DupyliCate (a) as tandems and proximals is comparable to that output by doubletrouble, while DupGen_finder gives a lower number of tandem and proximal duplicates. When comparing the total number of duplicates, DupyliCate and DupGen_finder have close estimates while doubletrouble gives a lower number of total duplicates. Singleton genes show

the maximum difference across the three tools. While DupyliCate and DupGen_finder output ~4000 and ~20 singletons respectively, doubletrouble does not have a singleton output. Further, when looking at the classifications assigned to all the input genes, both doubletrouble and DupGen_finder have remnant unclassified or unaccounted genes while DupyliCate classifies all the input genes.

The BUSCO-based threshold run used a self-normalized bit score cut-off of 0.696 based on the BUSCO-analysis and results in approximately two-thirds of the genes being classified as singletons.

Table 1: Consolidated results of benchmarking analysis on the *A. thaliana* Col-0 dataset. 'Misc.' gene duplicates in the table indicate all gene duplicates that are not tandem and proximal duplicates.

Tool	Tandem	Proximal	*Misc.	Singleton	Non-recurrant tandem and proximal duplicates	Non-recurrant total gene duplicates	Non-recurrant total duplicates and singletons	Total no. of coding genes input for classification	Total no. of unclassified genes
DupyliCate (a)	3584	1530	20092	3829	4844	23615	27444	27444	0
DupyliCate (b) (BUSCO-based threshold)	1754	707	5387	19909	2374	7535	27444	27444	0
doubletrouble	3045	3581	22224	-	4970	22869	22869	27444	4575
DupGen_finder	1316	2978	20921	19	4294	23340	23359	27444	4085

When compared to the other tools, DupyliCate additionally produces a duplication landscape plot from the distribution of the self-normalized bit scores in the genome of a sample analyzed for gene duplications (described in workflow). The plot of *A. thaliana* Columbia-0 (Col-0) is shown in Fig. 1.

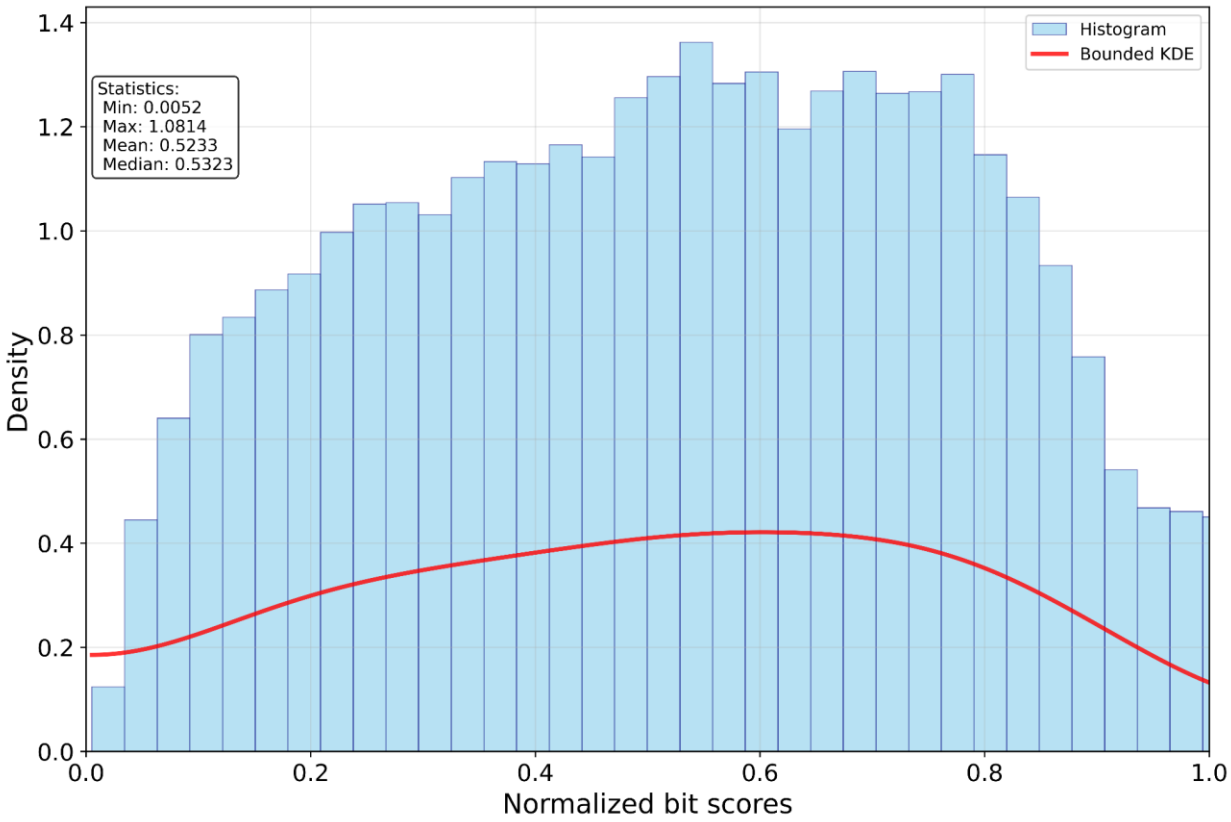

Figure 1: Duplication landscape plot of *A. thaliana* Col-0 output by DupyliCate - shows the distribution of self normalized bit scores. KDE stands for the Kernel Density Estimate.

In terms of runtime performance, doubletrouble and DupyliCate without BUSCO had comparable runtimes while DupGen_finder and DupyliCate with BUSCO had comparable run times (Fig. 2).

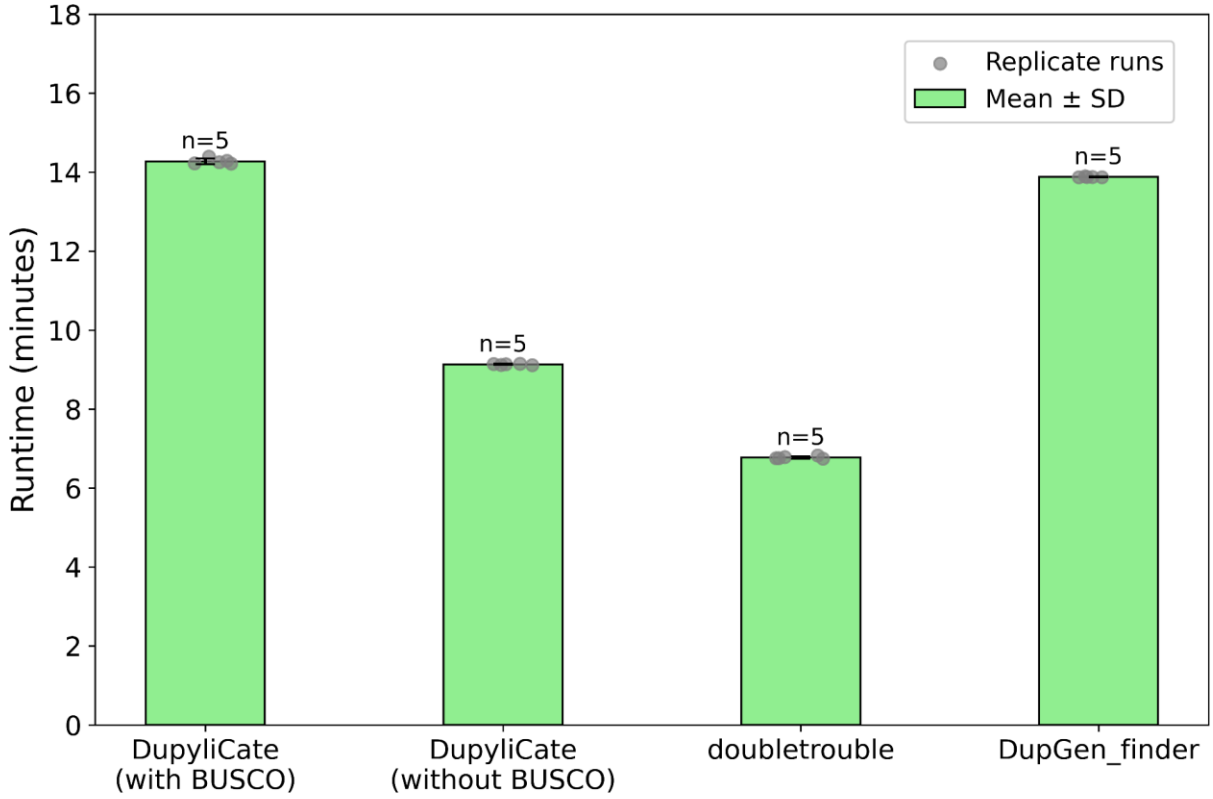

Figure 2: Run time comparison across DupyliCate, doubletrouble, and DupGen_finder

Run time analyses

The different parameters tunable in DupyliCate were explored in different runs of the tool using *A. thaliana* Col-0 dataset on an Ubuntu 24.04 LTS machine with 30 cores. For runs involving comparative analysis with a reference, *A. thaliana* Niederzenz-1 (Nd-1) was used as the sample and *A. thaliana* Col-0 was used as the reference. All the runs were performed using the default BUSCO-based thresholding method. The run times for the different runs are shown in Table 2 .

Table 2: Runtime compilation when using different modes and parameters of DupyliCate on *A. thaliana* datasets

Key parameter changed	Run time (mins)
Overlap mode	6.234
Strict mode	6.422
Phylogeny-based ortholog assignment in the presence of a reference	8.918
Local alignment-based ortholog assignment in the presence of a reference	8.0346
BLAST aligner	14.368
MMSeqs2 aligner	6.433
Expression analysis of all gene duplicates	71.82
Expression analysis of specific genes	28.317
Ka/Ks computation of all gene duplicates	25.5

The specific expression analysis of the *Xeroderma pigmentosum* (*XPB*), *β-L arabinofuranosidases* (*BLA*), and *Rapid Alkalinization Factor* (*RALF*) duplicates resulted in the matrix patterned expression plots shown in Fig. 3. The expression plot patterns for the *XPB* and *BLA* duplicates potentially point towards a subfunctionalization and neofunctionalization fate, respectively[17,18]. This is also supported by the correlation analyses performed among the gene duplicates (Supplementary_Table_S4). The *RALF* genes expression plot captures the pairwise expression of some genes in the *RALF* array[19]. The runtime trajectory analyses were performed using the Brassicales plant datasets used for the *FLS* case study. The runtime trajectory for duplicates mining (reference-free) (Fig. 4(a)), and duplicates mining combined with orthology assignments for comparative gene copy number analysis (with phylogeny) showed a linear time increase trend (Fig. 4(b)). The runtime for the large-scale study involving plant datasets from Phytozome with *A. thaliana* as the reference took 34.8 hours to complete. This demonstrates the suitability of DupyliCate for high throughput use cases involving gene duplicates mining as well as comparative genomic studies needing orthology assignments with respect to a reference.

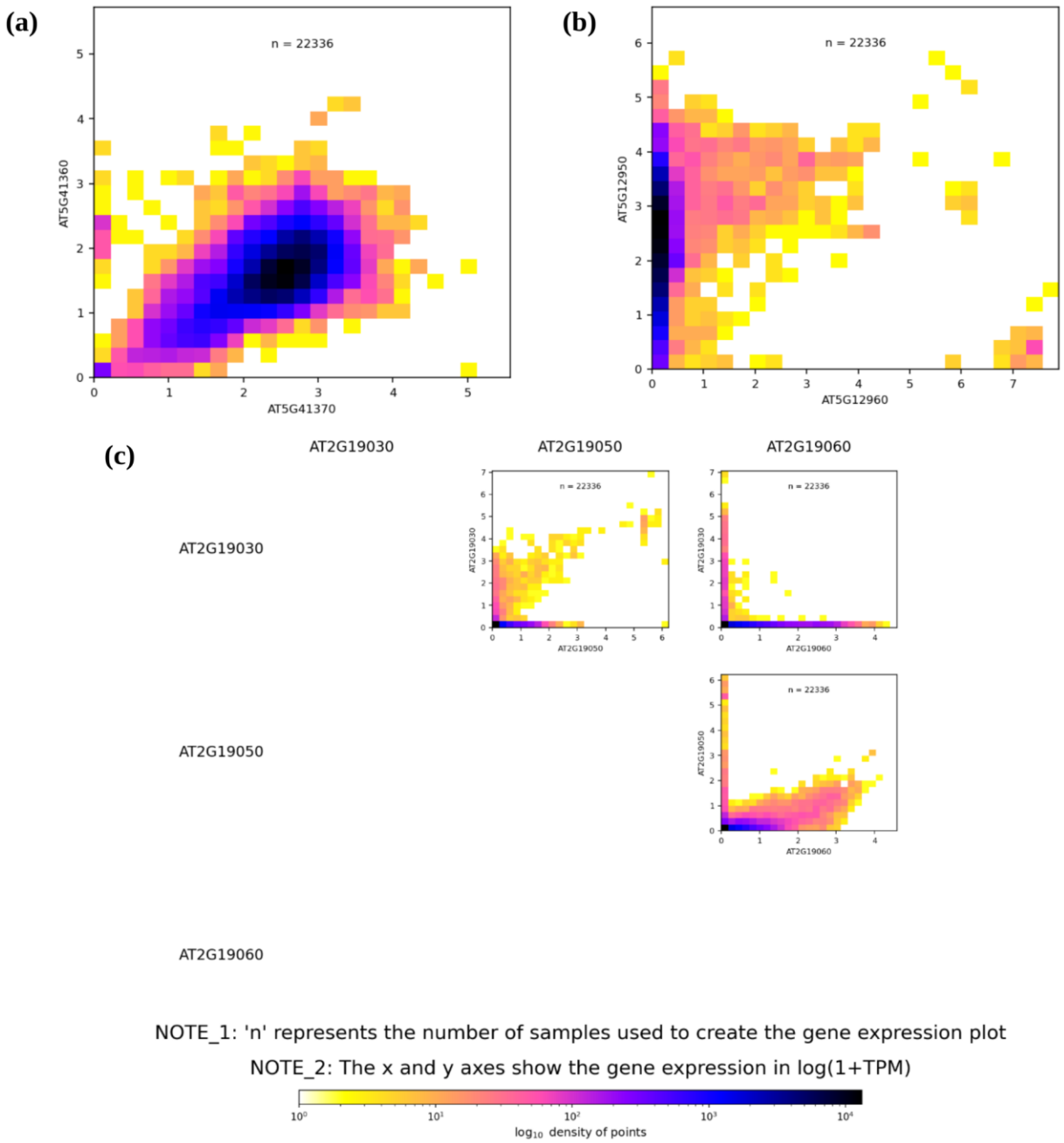

Figure 3: Pairwise gene expression plots of *A. thaliana* Col-0 tandem genes encoding (a) Xeroderma pigmentosum, (b) β-L arabinofuranosidases, and (c) Rapid Alkalinization Factors. TPM in the plot stands for Transcripts Per Million.

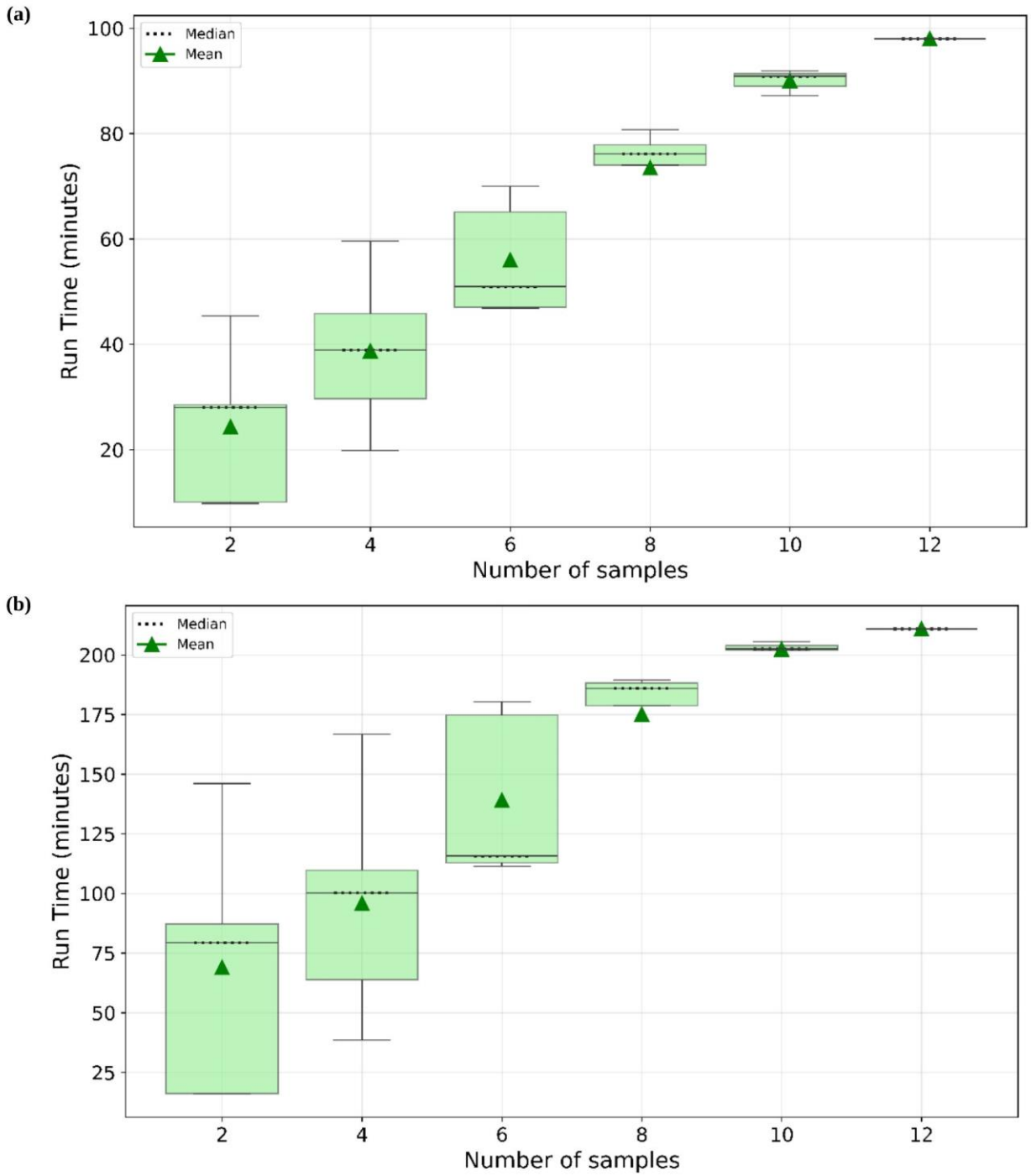

Figure 4: Run time trajectory analysis for (a) Gene duplicates mining across twelve Brassicales genomes (b) Gene duplicates mining, and comparative genomics involving ortholog assignments across twelve Brassicales genomes with respect to the reference *A. thaliana.*

Proof of Concept analyses

Proof of Concept (POC) analyses of DupyliCate were performed using well-known gene duplications reported in different plants. DupyliCate correctly identified tandem duplication of the *SEC10* gene in *A. thaliana* Nd-1 with respect to *A. thaliana* Col-0[20,21]. Next, tandem duplication of a gene in *Gardenia jasminoides* with respect to *Coffea canephora*[22] was correctly mined as a tandem duplicate array of four genes. A proximal duplication of genes was detected in the nematode tolerance locus of the nematode tolerant sugarbeet line StrubeU2Bv whereas the susceptible line KWSONT2320Bv had just a single copy of the gene[23]. DupyliCate confirmed the dispersed duplication of the sterol Δ^{24}-isomerase (24ISO) gene in *W. somnifera* responsible for withanolides biosynthesis[24]. These genes were also found to show up as copies along with the closely related sterol side chain reductase (SSR) genes in a proximal-dispersed gene duplicate group in the duplicate relationships file.

GeMoMa activating flag in the tool produced annotations for *V. amurensis* and *V. rotundifolia* with *V. vinifera* as the reference, and helped identify the gene duplicates for species with missing annotation files. *V. amurensis* had the highest number of gene duplicates (Table 3). Among the analyzed rice and related genomes, *Echinochloa colona* and *Echinochloa oryzicola, which* are rice weeds, had the highest number of gene duplicates. All the other rice varieties and their relatives had a comparative number of duplicates (Table 3). The copy number table (Supplementary_Table_S5) gave a visual presence-absence matrix of the orthologs of the different stress genes in *O. sativa* indica across the analyzed rice relatives.

DupyliCate also identified gene duplicates and singletons in the non-plant species - *E. coli, S. cerevisiae and C. elegans*. As expected, *E. coli* and *S. cerevisiae* had a very low number of gene duplicates whereas *C. elegans* showed a relatively high number of gene duplicates (Table 3).

Table 3: Consolidated results of gene duplicates identified and classified in *Vitis* species, *Oryza* species, rice relatives, and non-plant samples

Samples	Tandem duplicates	Proximal duplicates	Dispersed duplicates	Singletons	Low confidence duplicate groups	Moderate confidence duplicate groups	High confidence duplicate groups
V. vinifera	4856	2095	8751	10832	98	26	5867
V. rotundifolia	5281	2958	10829	10898	58	23	6849
V. amurensis	5422	3741	17364	10463	83	24	9632
L. perrieri	1336	1006	5883	21241	47	7	3587
O. barthii	1518	1470	6623	25389	43	9	4254
O. glaberrima	1631	922	9641	21406	47	11	5065

O. nivara	1282	1269	7253	26941	47	9	4326
O. rufipogon	1252	1342	7018	27885	56	9	4237
O. sativa japonica cv.	1788	1686	7300	25582	42	9	4676
Z. latifolia	1033	724	17702	24747	169	8	8317
E. oryzicola	5428	2933	53451	9743	921	275	24672
E. colona	5922	2798	77576	10854	1347	362	28510
E. coli	29	22	570	3673	0	0	267
S. cerevisiae	101	42	971	4936	12	1	468
C. elegans	2277	978	3609	13727	5	41	2806

FLS evolution in Brassicales

Among the Brassicales plants analyzed in the *FLS* case study*, I. tinctoria, E. vesicaria, S. alba* and *B. napus* showed a high number of gene duplications. Tables consolidating the gene duplicates mined and the orthologs and copy numbers of the *Arabidopsis thaliana FLS* (*AtFLS*) genes in the other Brassicales plants are included in Supplementary_Table_S3 and Supplementary_Table_S6 respectively. The presence-absence variation in the copy number table showed the presence of *FLS1*- the functional *FLS* in all the Brassicales plants analyzed. But all of the Brassicales plants were missing at least one or more of the other *FLS* genes. The phylogenetic tree used to validate the ortholog assignments is shown in Fig. 5. Most of the ortholog assignments in DupyliCate and the clade grouping in the phylogenetic tree match, except for a few sequences (coloured with red strips in the tree (Fig. 5). The *C. papaya* gene identified as *FLS1* (evm.TU.supercontig_37.77) does not exactly group in the *AtFLS1* clade but is a sister to this clade. No other FLS genes were identified in *C. papaya.* The *B. oleracea* gene that groups in the *AtFLS2* clade (Bol019362.v1.0) shows up as an FLS1 ortholog in DupyliCate. On a similar note, a group of genes marked by the red strips in the *FLS4* clade in the tree are assigned as orthologs of *FLS1* or *FLS3* in DupyliCate results. A consistent occurrence of this pattern among these genes that form a separate subclade with respect to *AtFLS4* is interesting. The phylogenetic tree also shows a high number of sequences in the *FLS3* and *FLS4* clades, indicating a potential expansion of these genes in Brassicales.

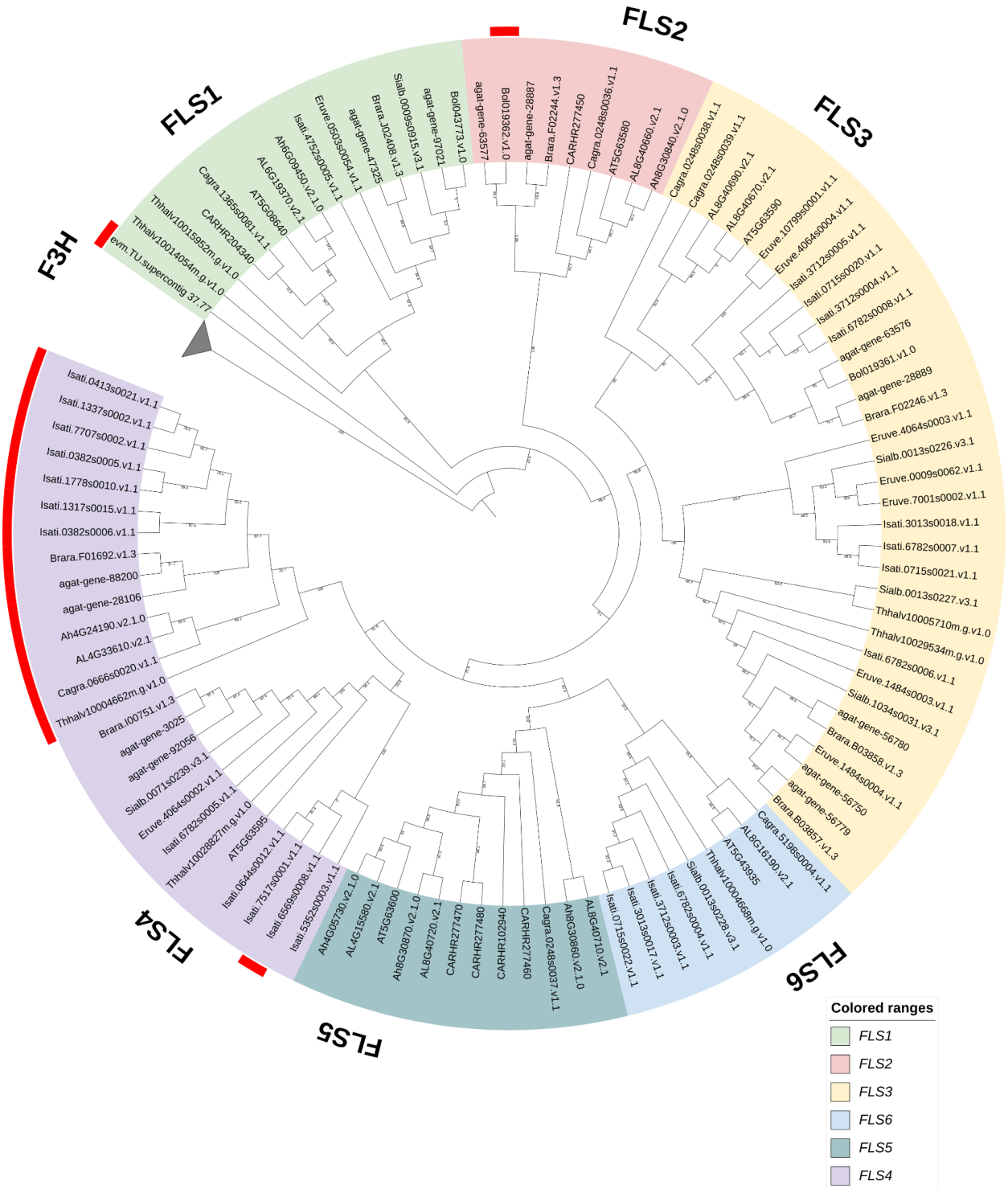

Figure 5: Phylogenetic analysis of *FLS* genes across twelve Brassicales plant genomes. The red striped sequences have different ortholog assignments in DupyliCate and the phylogenetic tree. The collapsed clade contains flavanone 3-hydroxylase (F3H) sequences as outgroups

MYB evolution across a broad phylogenetic range of plant species

A large-scale analysis of plant datasets from Phytozome facilitated the understanding of *AtMYB12,* and *AtMYB111* evolution across a wide range of plant species. A summary of the number of gene duplicates mined across the 153 plant species along with the copy number table representing the *A. thaliana MYB12*, and *MYB111* orthologs across the analyzed datasets were obtained (Supplementary_Table_S7). Among the analyzed plants, orthologs for *MYB12* and *MYB111* were not found in algae, mosses, liverworts, and ferns. Interestingly, the *MYB12,* and *MYB111* orthologs were found in lycophytes (club mosses). In the other plants, orthologs were found for at least one or both of *AtMYB12* and *AtMYB111*. The phylogenetic tree used to validate the ortholog assignments is shown in Fig. 6. The phylogenetic analysis also reveals an interesting evolutionary pattern of *MYB12* and *MYB111*. As seen in Fig. 6, *AtMYB12* and *AtMYB111* fall in separate clades that consist of sequences from *A. thaliana* and other dicot plants (pink and purple clades in Fig. 6). Monocots, a few other dicots and some basal angiosperms on the other hand do not show separate lineages of these flavonol biosynthesis-regulating MYBs.

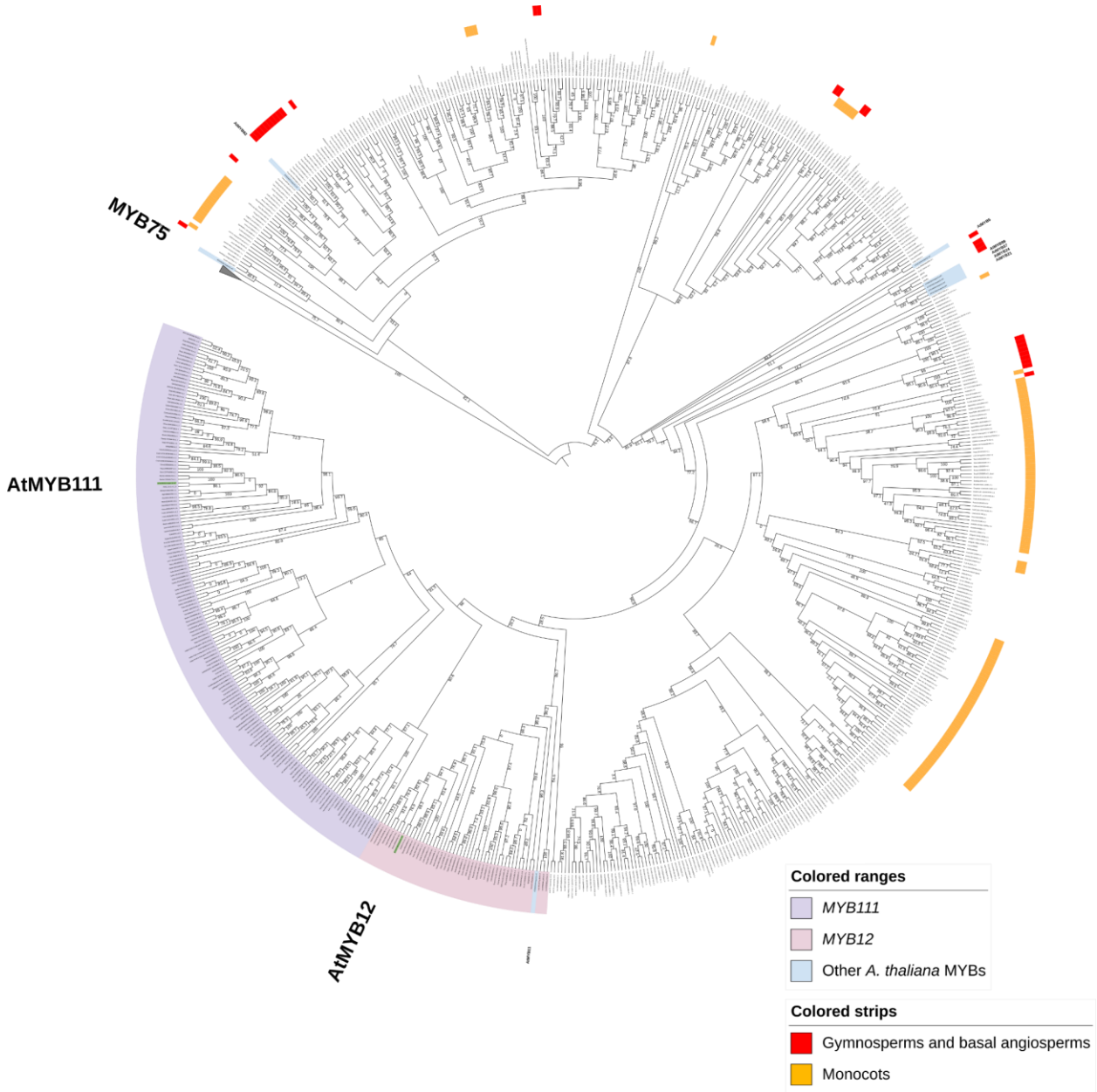

Figure 6: Phylogenetic analysis of *MYB12* and *MYB111* genes across a broad phylogenetic range of plants. All the sequences not covered by the yellow strips belong to dicots. The collapsed clade contains *MYB75* sequences as outgroups.

Further, the tree revealed that there are *MYB111* and *MYB12* orthologs captured correctly, as well as false positive sequences (uncoloured sequences bounded by the outgroup MYB75 clade and the *A. thaliana* flavonol MYB clade coloured in blue). The ortholog assignments in DupyliCate and the clade grouping in the tree were compared (Fig. 7).

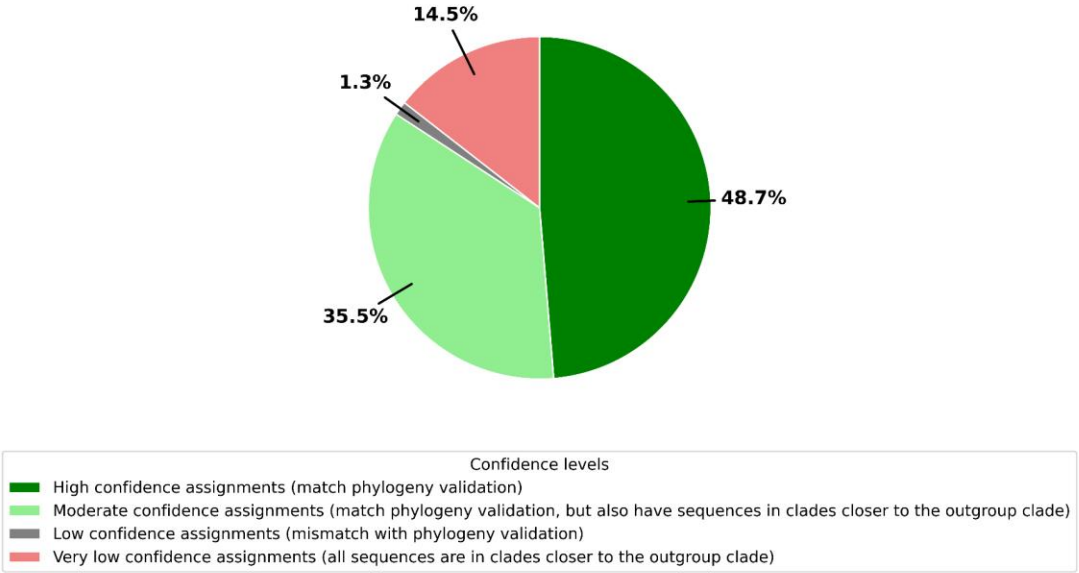

Figure 7: Pie chart representing the validation of DupyliCate's ortholog assignments with respect to clade grouping in the phylogenetic analysis

The validation analysis shows that nearly 50% of the ortholog assignments in DupyliCate matched exactly with the phylogenetic tree. More details on the validation analysis can be found in Supplementary_Methods.

Discussion

The benchmarking analyses reveal that DupyliCate performs on par or better than the other tools doubletrouble[15], and DupGen_finder[9] when using a manual threshold specification. This can be seen from the remnant number of unclassified genes in the other two tools (Table 1). doubletrouble does not specify or produce a singleton output file, hence it can be explained that it inherently considers the genes not assigned to the duplicate groups as singletons. DupGen_finder on the other hand, outputs only those genes that do not have self local alignment hits as singletons. This ignores genes that hit up only against themselves in local self alignment. In the manual threshold specification to DupyliCate, genes that do not have self hits, genes that have only themselves as self hits, and genes whose second best hit within the genome fail to pass the threshold are classified as singleton genes, resulting in a few thousand singleton genes. But when using a BUSCO-based threshold for obtaining gene duplicates, genes that do not have self hits, genes that have only themselves as self hits, and genes whose second best hit within the genome fail to pass the BUSCO-derived threshold are classified as singleton genes. This causes the results to significantly differ giving around 27.5 % of gene duplicates in *A. thaliana*. However, the relatively lower number of gene duplicates in *A. thaliana*

is supported by a low BUSCO duplication percentage (Supplementary_Table_S2). A potential biological explanation could be that only a certain proportion of genes were retained as duplicates after the polyploidization events reported in *A. thaliana*[25]. These estimates are also close to the estimate of 28% gene duplicates in *A. thaliana* previously reported[26] after the most recent alpha polyploidy event (Supplementary_Figure_1)[25,27]. This puts forth the biological implication of the BUSCO-based thresholding method that distinguishes genes that escaped ancient duplication events and those that were retained in gene families or underwent more recent duplications. It thus allows for species-specific gene duplicates identification and classification. But it is also important to note that there is no clear boundary between singleton genes and duplicated genes. Both are rather extremes on a gradient. Therefore, the identification is imperfect, influenced by the cutoffs and other selection criteria determined by the users. Further the different number of singleton gene outputs in the different existing tools could as well be attributed to the fluidity of the singleton-duplicate segregation criteria. However, the BUSCO-based approach developed in this study offers one possible way of segregating singleton and duplicated genes with a reliable biological support. Additionally, the duplication landscape plot of *A. thaliana* output by DupyliCate (Fig. 1) also offers a visual inference on the genome-wide duplication status. Left and right skewed distributions in the plot could suggest genomes with extremely low gene duplications (as seen for *E. coli, S.cerevisiae* (Supplementary_Results_1)) and genomes with extensive gene duplications or polyploidization (Supplementary_Results_1), respectively. The approximately central, bell-shaped curve for *A. thaliana* (Fig. 1) corresponds well with the polyploidization-gene duplications loss events described for *A. thaliana*[26,28]. Looking at the performance metrics of the three tools (Fig. 2), DupGen_finder had the highest run time, while doubletrouble had the lowest run time. However, since an outgroup was specified in DupGen_finder, the inclusion of a forward alignment step leads to an increase in run time. Thus doubletrouble and DupGen_finder can be considered to be on par with respect to run times. Since DupyliCate incorporates FASTA cleaning, GFF validation, and duplicate group clustering steps, the runtime is slightly longer, especially when BUSCO analysis is included. Also steps like isoform removal and execution of local alignment, which need to be performed separately by the user prior to running the other tools doubletrouble, and DupGen_finder, are integrated in DupyliCate's run. Although this increases the runtime, it could lead to possibly reduced hands-on time when using DupyliCate. Moreover, it is expected that the availability of DupyliCate via docker and conda would further simplify the installation and use of the tool, and expand its usability for users with different levels of bioinformatics expertise.

The run time analyses involving different parameter and mode selection with the *Arabidopsis* datasets give an idea of the runtimes for the different functionalities. As seen from Table 2, a complete expression analysis, and Ka/Ks analysis of gene duplicates take longer. Hence, when such analyses are intended, it is recommended to use as many cores as available to complete the run in the shortest possible time. The expression analysis plots output by the use of the specific duplicate analysis flag in the absence of a reference (Fig. 3) help analyze the functional fates of the duplicated genes and determine expression divergence among them. The matrix

pattern of the plot helps visualize the gene expression trends among all the genes in the same duplicate group at once. The entire RALF tandem duplicates array reported by Cao and Shi[19] - *AT2G19020, AT2G19030, AT2G19040, AT2G19045,* was identified in the RALF tandem group by DupyliCate along with the genes *AT2G19050,* and *AT2G19060.* However, *AT2G19020, AT2G19040, AT2G19045* were found to be perceived (putative) pseudogenes based on the expression analysis. Thus Fig. 3(c) shows the pairwise expression plots of the other genes in the identified tandem duplicates array. But it is important to note that including information on the spatial aspects of gene expression in the different tissues and under different conditions is necessary to confidently ascertain the subfunctionalization, neofunctionalization or pseudogenization fates of the gene duplicates. From Fig. 4(b) it is evident that including phylogeny in the presence of a reference increases the run time in combination with the dataset size. But incorporating phylogeny during ortholog assignments will produce better ortholog assignments, as local alignment-based approaches alone are not enough to identify orthologs correctly[6]. Hence phylogeny is recommended when a reference is used. In case if the user just wants to have a first hand assessment of gene copy numbers with comparative analysis, and accuracy is not important, then the phylogeny option can be deactivated.

The *SEC10* gene is incorrectly collapsed in the Col-0 sequence, but is properly represented in *A. thaliana* Nd-1's assembly and annotation[20,21,29]. This was confirmed through the DupyliCate analysis where the *SEC10* gene duplication was identified as a tandem duplication in the Nd-1 dataset. The first dedicated gene in the crocine biosynthesis pathway of *Gardenia jasminoides* is a part of a four-gene tandem array that shares just one ortholog with the *Coffea canephora* genome. This tandem duplication-led expansion of the gene cluster is responsible for the initial critical steps in crocine biosynthesis that is not found in coffee[22]. The entire tandem cluster was correctly identified by DupyliCate demonstrating the usefulness of representing gene duplicates in cluster groups instead of gene pairs. In the analysis of nematode resistance locus between tolerant and susceptible sugarbeet lines[23], DupyliCate identified that the crucial genes from the locus belong to a proximal duplicate cluster. This shows that proximal duplication was responsible for the nematode tolerance phenotype seen in the tolerant line StrubeU2Bv while exemplifying the application of DupyliCate to answer biological questions. Finally, the 24ISO gene responsible for the first steps of withanolide biosynthesis in *W. somnifera* was identified to be present in a dispersed duplicate group. It was also found to appear in the duplicates relationships results along with the closely related SSR genes[24] providing a PoC for the usefulness of the duplicates relationships file in the overlap mode and the mixed duplicates file in the strict mode. These different PoC analyses show the ability of DupyliCate to correctly identify gene duplicates and classify them into the different duplication classes. The gene duplicates identification in the *Vitis* species lacking annotation was facilitated through the GeMoMa flag of DupyliCate (Table 4). This option could be useful in cases where the samples of interest lack structural annotation or a good quality structural annotation, but have a closely related species with a good structural annotation that can serve as the hint for gene prediction with GeMoMa. It is also helpful to note that the hint for GeMoMa annotation and the reference for comparative analysis with DupyliCate could be the same or different. This means, the GeMoMa flag can be used in the presence as well as in the absence of a reference for

DupyliCate analysis, but requires the specification of a hint for creating the annotation for the sample. In either case, the species whose annotation is to be used by GeMoMa for the new annotation must be specified in the GeMoMa config TXT file along with the names of the samples to be annotated as well. More details are provided in the GitHub repository. In the rice genome sequence analysis, the rice weeds showed a high number of gene duplicates (Table 4). The BUSCO analyses (Supplementary_Table_S2) showed that these rice weeds have a high percentage of duplicated BUSCO genes (>80%), and the pseudo-ploidy numbers (apparent indicators of ploidy of the sample based on BUSCO gene frequencies) of *E. colona* and *E. oryzicola* were ~3 and ~2, respectively. Their duplication landscape plots (Supplementary_Results_1) were also distinctively right skewed suggesting a recent polyploidization event. The other rice varieties had a comparable number of gene duplicates and showed a very similar trend of duplication landscape plots except *Oryza glaberrima.* The plot of *O*. *glaberrima* was almost flat except for a marked high in the number of genes with a self-normalized bit score of 1 indicating that the duplicates in the genome are recent or have not diverged greatly (Supplementary_Results_1). DupyliCate's applicability to non-plant samples was shown by mining duplicate genes in *E. coli, S. cerevisiae*, and *C. elegans*. The high number of gene duplicates identified in *C. elegans* corroborates the widespread occurrence of gene duplications in *C. elegans*, when compared to other non-plant eukaryotes like yeast and fruit fly[30].

Among the analyzed Brassicales plants for the *FLS* case study, *I. tinctoria, E. vesicaria*, and *S. alba* had a pseudo-ploidy number of ~2 indicating their diploidy, while *B. napus* with a pseudo-ploidy of ~3 was classified as triploid (Supplementary_Table_S2). This explains their high gene duplication numbers. Coming to the ortholog assignments in the copy number analysis, *C. papaya FLS1* (evm.TU.supercontig_37.77) groups as a sister of the *AtFLS1* clade and is the only *FLS* ortholog identified. It also has a good BUSCO completeness (>80%) (Supplementary_Table_S2) indicating a biological phenomenon. *C. papaya* belongs to Caricaceae, one of the most basal families in Brassicales[31]. This along with the phylogenetic tree (Fig. 5) suggests a possible hypothesis on *FLS* evolution in the Brassicales - the true *FLS* had appeared in the early Brassicales plants followed by an expansion of the *FLS* genes in Brassicaceae after the divergence of Caricaceae and Brassicaceae. The *B. oleracea* gene showing up in the *AtFLS2* clade (Bol019362.v1.0) could probably be an ancient *FLS2* duplicate that cannot be placed correctly only based on local alignment methods. Similarly, almost all the sequences falling in the later subclade of *AtFLS4* (coloured by the red strips in *AtFLS4* clade in Fig. 5), have *AtFLS4* appear as only the fourth or the fifth best forward hits of these genes in the forward local alignment. These genes could also be ancient duplications of *FLS4* and hence difficult to assign orthologs based on candidates from similarity based methods. Such difficult cases of ortholog detection can be resolved with higher phylogenetic resolution by including more sequences that give better placement and context to the query gene within its gene family supported by better taxon sampling and global alignment cues[32]. Further, from the phylogenetic analysis, there is an evident expansion of *FLS3*, and *FLS4* genes happening in Brassicaceae, which is quite intriguing. This is also observed in the *FLS* phylogenetic tree presented by

Schilbert and co-workers[33]. Expression analyses performed in DupyliCate, found quite a majority of genes in the *FLS3*/*5* clades to be pseudogenes. In the *FLS4* clade most of the *I. tinctoria* genes turned out to be pseudogenes. A potential explanation for the *FLS3/4* expansion could come from the expression analysis. Since these are non-functional *FLS* genes, they could be under relaxed selection pressure. Expansion might be tolerated because additional copies do not impose strong fitness costs, allowing them to accumulate mutations and gradually become pseudogenes.

The transcriptional regulation of flavonol biosynthesis in plants is achieved by the action of *MYB11, MYB12*, and *MYB111* transcription factors (TFs)[34]. Schilbert and Glover analyzed the evolution of flavonol regulators in the family Brassicaceae and reported the duplication of *MYB111* and *MYB12* TFs in the family[35]. The MYB case study presented here is presented as a study of these flavonol regulators' evolution across a large phylogenetic range of plants. Flavonoid biosynthesis is well known and established in land plants. Flavonols, a class of flavonoids that play a major role in protecting against abiotic stress factors like UV exposure are also largely found in land plants[36–38]. These factors provide a basis for the reason behind the absence of flavonol regulating *MYB12* and *MYB111* orthologs in non-land plants and basal plants like mosses. The *MYB12,* and *MYB111* orthologs mined in club mosses were examined and were found to belong to very low confidence ortholog groups, which could mean that these species could also lack the *MYB12,* and *MYB111* orthologs. An alternative explanation could also be that algae, mosses and other basal plants might have other flavonol biosynthesis regulating MYBs. Looking at the other plants in the analyses, all assignments that had very high confidence orthologous group confidence scores (OGCS) in DupyliCate were found to match exactly with the phylogeny results. While groups with moderate and low OGCS were found to have mismatches with the tree or have false positive sequences. This demonstrates the usefulness of the OGCS scoring system developed that can help refine and obtain more reliable results. Further, nearly half of the examined plant species' ortholog assignments are exact matches with the tree with nearly one third having matches with the presence of false positive sequences (Fig. 7). One possible reason for the ortholog misassignments observed could be the widely varying evolutionary distance differences between the reference *A. thaliana* and the analyzed sample species. This is also supported in the phylogenetic tree where most of the sequences near the outgroup clade belong to either gymnosperms, basal angiosperms, monocots, or polyploid plants like cotton. The phylogenetic tree (Fig. 6) also provides some biological insights. It indicates the duplication of *MYB12* and *MYB111* in some dicots (purple and pink clades) that could have been derived from an ancestral flavonol regulating MYB which is still observed in gymnosperms, basal angiosperms, monocots, and some dicots. In conclusion, the MYB case study demonstrates the usability of the tool for large scale analysis involving gene duplicates mining, gene copy number analysis in nominal time and its ability to provide biological insights.

Further, a detailed discussion on parameter tuning of DupyliCate is provided in Supplementary_Discussion.

In conclusion, DupyliCate, a high throughput Python script has been developed for gene duplicates identification, classification and integrated expression analysis. It has a number of adjustable parameters, species-specific singleton duplicates segregation, flexibility to handle a variety of GFF files, synteny analysis of small scale duplicates (in the presence of a reference) and optional phylogeny-based ortholog detection for comparative gene copy number analysis. DupyliCate can thus be a valuable addition to the bioinformatics toolkit to support research in the domains of gene duplications research and comparative genomics for both plant and non-plant species.

Materials and methods

Datasets

Datasets of *Arabidopsis thaliana* Col-0 and Nd-1[21] were used to demonstrate the available modes, different aligner options, possible Ka/Ks analysis and expression analysis functionalities of DupyliCate. The two accessions were also used to demonstrate the PoC for *SEC10* duplication in the *A. thaliana* Col-0 dataset. *Gardenia jasminoides* and *Coffea canephora* datasets[22] were used to demonstrate the PoC for tandem duplication of the first gene in crocine biosynthesis in *G. jasminoides* with respect to *C. canephora.* Datasets of the *Beta vulgaris* lines StrubeU2Bv and KWSONT2320Bv[23] were used as PoC for proximal duplication of the nematode tolerance gene in the tolerant line StrubeU2Bv with respect to KWSONT2320Bv. Datasets of *Withania somnifera* and *Physalis pruinosa*[24] were included to show PoC for dispersed duplication of the 24-Isomerase gene in *W. somnifera* with respect to *P. pruinosa*. DupyliCate was used on three *Vitis* datasets (*Vitis vinifera, Vitis rotundifolia,* and *Vitis amurensis*) to demonstrate applicability on complex genome sequences and also to demonstrate the use case of GeMoMa feature of the tool using the datasets of *V. rotundifolia* and *V. amurensis* that lacked structural annotation. A collection of 10 genome sequences of rice (*Oryza sativa*) and close relatives from the RiceRelativesGD database[39] were leveraged to demonstrate the performance of DupyliCate on monocot plants. Gene duplicates were mined in *Oryza sativa indica* variety and its relatives. The stress related genes in *O. sativa indica* sp. were used to create a copy number table to analyse the number of the corresponding orthologs in the other rice relatives. *Escherichia coli*[40]*, Saccharomyces cerevisiae*[41], and *Caenorhabditis elegans*[42] datasets were used to show PoC for deployment on non-plant species.

Application case studies

Arabidopsis thaliana is reported to have six copies of the flavonol synthase (*FLS1-6*) gene encoding a key enzyme in flavonoid biosynthesis[43]. Of the copies, only *AtFLS1* is known to produce a catalytically functional protein whereas others could be potential pseudogenes whose function is yet to be uncovered fully. Comparative analyses of the FLS genes in the Brassicales

order combined with expression analysis can help better understand the evolutionary dynamics of this gene. A total of 12 datasets of the Brassicales family (detailed in Supplementary_Table_S1), for which expression data were available, were chosen to do an application case study on flavonol synthase evolution in the Brassicales using *Arabidopsis thaliana FLS1-6* as reference genes.

Next, a large collection of 153 plant datasets was retrieved from phytozome (details specified in Supplementary_Table_S1). These datasets were used to undertake a large-scale gene duplicates mining study across plant datasets to demonstrate the high throughput utility of the tool. Further, these datasets were also used to study the evolution of the key flavonol regulator MYBs (Myeloblastosis genes) - *MYB12* and *MYB111* across the plant kingdom with *Arabidopsis thaliana MYB12* and *MYB111* as reference genes, to demonstrate the biological use case. The forward normalized bit score and forward similarity cutoffs were lowered to 0.2 and 30%, respectively, since the run involved samples quite distant from the reference *A. thaliana* to avoid missing potential orthologs in the samples.

The results of DupyliCate from the two application case studies were also validated by building a phylogenetic tree. For the FLS tree, flavanone 3-hydroxylase (F3H) sequences were used as outgroups. For the MYB tree the ortholog results from DupyliCate were visualized in a phylogenetic tree with anthocyanin regulating *MYB75* sequences as outgroups. Other closely related *A. thaliana* MYBs *AtMYB5, AtMYB11, AtMYB21, AtMYB24, AtMYB57, AtMYB99, AtMYB82, AtMYB123* were also included in the tree. MAFFTv7.5.2.6[44] was used with the G-INS-i accuracy-oriented method for alignment of the amino acid sequences. pxclsq from PHYX[45] was used to remove alignment columns with occupancy below 10%. Phylogenetic trees were inferred using IQ-TREE3[46] with the parameters -m MFP -wsr and --alrt 1000 -B 1000[47]. The best-fit evolutionary model inferred with ModelFinder[48] for the FLS case study was Q.PLANT+R5 and for the MYB case study was JTT+R8. The phylogenetic trees were visualized using iTOL v7.2.2[49].

All the datasets were analyzed with DuypliCate v1.0 in the overlap mode with BUSCO-specific threshold option activated. In case, a reference was used, phylogeny was used to determine the ortholog relationships with the reference genes.

Workflow

A pipeline for gene duplicates identification, classification, and further downstream analyses was implemented in Python3 (Fig. 8). The overall workflow involves validation of input files followed by self alignment of the sample(s) (sample(s) - species in which gene duplicates are to be identified). Species-specific thresholds are then used to segregate singleton genes from duplicates and duplicates are categorized as tandem, proximal, dispersed and mixed (a class where genes can belong to more than one duplication class; this helps retain the relationship

across the different duplicate groups). In the presence of a reference organism (reference - the species with respect to which the sample's genes are assigned orthologs and synteny analysis is performed), forward alignment is performed. The classified gene duplicates are assigned to orthologs in the reference, and synteny analysis is performed for small scale duplicate groups of tandem duplicates and proximal duplicates. Depending on the mode chosen, a fourth category of mixed duplicates is returned. DupyliCate also facilitates optional gene expression analysis of gene duplicates. Further, it provides the user with the option to calculate Ka/Ks values of the duplicates.

The DupyliCate script was tested with its dependencies and tools whose version numbers are specified in the workflow steps detailed below.

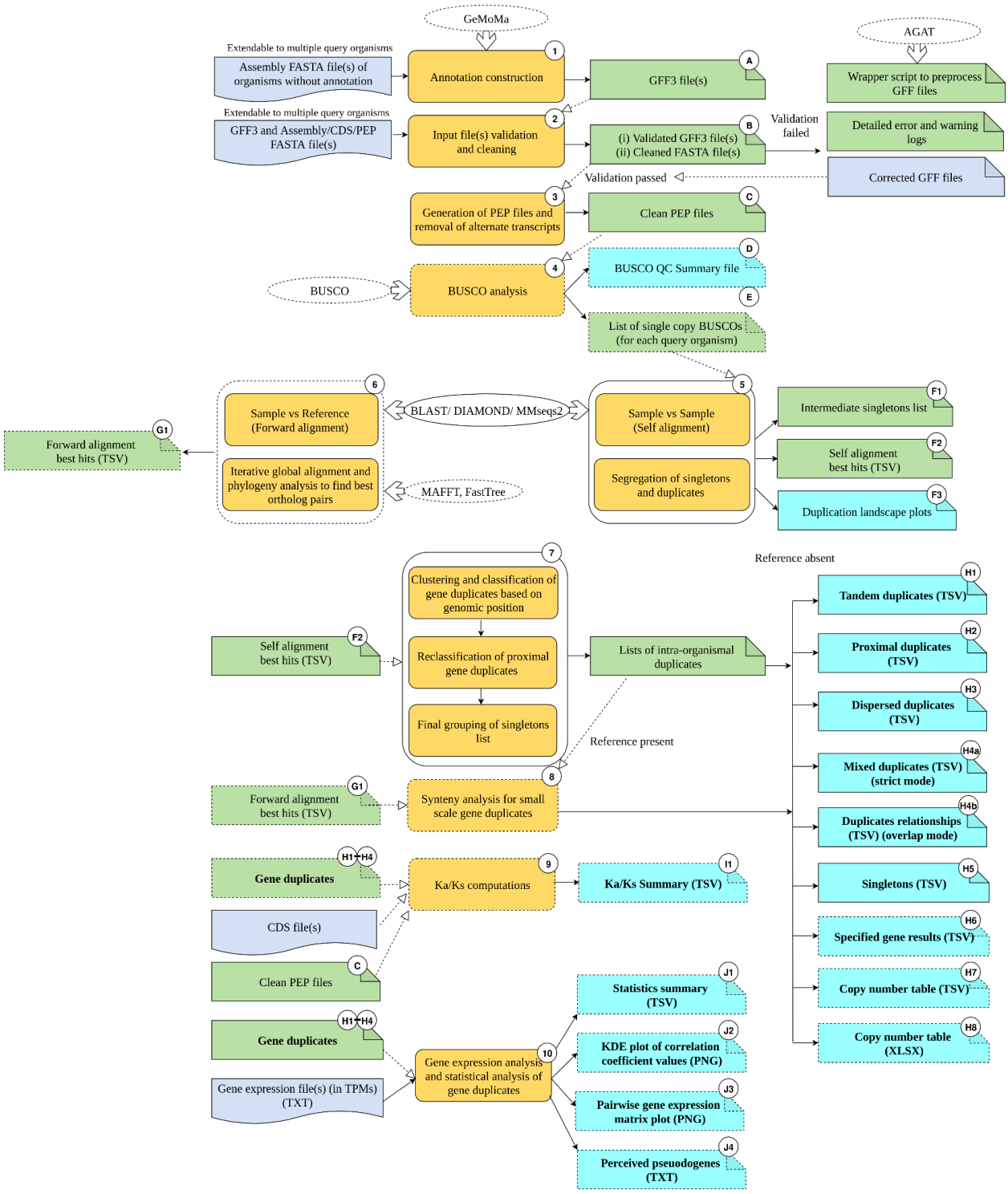

Figure 8: Schematic representation of DupyliCate's workflow. The steps of the pipeline are numbered and the output files are alphanumerically labelled. The blue boxes indicate user provided input files and green boxes indicate intermediate output files produced by the pipeline that serve as inputs for further steps. Pipeline steps are colored yellow. Output files are shown

by the cyan coloured boxes. External tools are represented by the ovals. Dotted outlines around the shapes indicate that they are not a part of the mandatory inputs, and dependencies or routine outputs of the pipeline.

Step 1: Optional structural annotation: DupyliCate requires the structural annotation file(s) (GFF3) along with one of the following (1) assembly sequence, or (2) coding sequences, or (3) peptide sequences in FASTA format. In case structural annotation is not available for a sample in which duplicates need to be analyzed, there is an option to provide the annotation of a related species as a hint for the gene prediction process with GeMoMa v1.9[50].

Step 2: Input file(s) check and validation: The input files are first checked and validated before the actual run starts. If the check fails, the script exits and the errors are recorded and displayed sample-wise. In case of GFF errors, there is a helper script (AGAT_wrapper.py) provided along with the main script that uses AGAT v1.5.1[51] to process and correct the GFF files.

Step 3: GFF-FASTA files standardization, error correction and peptide file generation: GFF files have widely varying attribute fields denoting the information about the different genomic features. It is also necessary for the headers in the protein FASTA file to match correctly with the text denoted by these different attribute fields. This is a major limitation hampering the deployment of a number of existing tools on different GFF files. To address this issue, this work provides a two-step solution. First, it offers another helper script (FASTA_fix.py) for the users to process the FASTA file headers according to the attribute field text in the respective GFF files. Next, the main DupyliCate script offers an option to the user to indicate the specific GFF attributes to be used by the script when processing the GFF file of a particular species, through a simple config TXT file. This helps to circumvent GFF-FASTA mismatch errors, and expand the usability of DupyliCate on a number of GFF file formats. The corrected and validated files then enter the main analysis and are processed to generate the protein FASTA files without alternate transcripts.

Step 4: Quality check of input peptide files: Since the output depends on the quality of the input files and is also influenced by the ploidy of the samples being analyzed, a BUSCO[52] - based QC step is included. This provides detailed QC reports containing the BUSCO completeness, duplication, and BUSCO-derived pseudo-ploidy number that indicates the ploidy of the input sample(s).

Step 5: Threshold determination for singleton-duplicate segregation and self alignment: Before moving on to the duplication analysis, it is important to segregate singletons and duplicates correctly. This in turn depends on the duplication landscape within each analyzed species. A new method for such species-specific threshold determination based on BUSCO v6.0.0 is implemented here. For this, two cut-offs - one based on self normalized bit score (bit score of the hit of a polypeptide sequence with another polypeptide sequence within the genome (second best hit) relative to the bit score against itself) and self-similarity are offered. The complete BUSCO single copy genes are retrieved for each sample. If enough BUSCO single copy genes are present, the normalized bit score of the 95th percentile of the BUSCO single copy genes is picked as the threshold for duplicates and singletons identification. By default, this BUSCO-based auto threshold method is chosen. If BUSCO is not available or if the

sample has a very low number of BUSCO single copy genes, then the fallback is to go for a self-similarity threshold instead of normalized bit score threshold. There is also an option to manually set the normalized bit score threshold and self-similarity threshold, allowing for fine grained analysis. The detailed steps of threshold determination can be found in Supplementary_Figure_2.

Next, self alignment of the sequences of a sample is performed. Options are provided to choose between DIAMOND v2.1.14[53] BLAST v2.17.0[54], and MMseqs2[55] for the alignment step. Following self-alignment, duplicates and singletons are segregated using the previously determined threshold. Independent of the thresholding method, a plot folder called Duplication_landscape_plots is produced. It contains the normalized self-bit score distribution plots of samples. The overall pattern of the duplication landscape plot can be used to infer the genome-wide duplication status of the species.

Step 6: Ortholog determination: If sequences of a reference are provided, forward alignment of the sequences of each sample is performed against the sequences of this reference. Here again, two thresholds based on similarity value of the hits in the forward alignment and forward normalized bit score (bit score of a forward alignment hit of a polypeptide sequence relative to the bit score of the hit of the polypeptide sequence against itself in self alignment) are applied to retain good quality forward hits for the further downstream steps. Next follows a comprehensive ortholog detection step in the presence of a reference. The top n candidates (default value of n=10) from local alignment algorithms like DIAMOND are subjected to a global alignment using MAFFT v7.5.2.6[44], synteny analysis (in-built code), and gene tree inference using FastTree2[56]. Phylogeny derived evolutionary distance assessments using DendroPy[57] are utilized to obtain orthologs of genes in the samples among the reference genes. The detailed steps in ortholog detection are shown in Supplementary_Figure_3. Since a multi-evidence approach is used for ortholog search, a confidence scoring scheme was developed to ascertain the confidence of each of the assigned orthologs based on the number of evidence points needed to determine orthology. The ortholog confidence scoring scheme is detailed in Supplementary_Table_S8.

Step 7: Gene duplicates grouping and classification: Next, the duplicates clustering and grouping step is performed per sample to give rise to the different gene duplicate groups shown in Fig. 9 below. There are two main modes of DupyliCate. Gene duplicates are classified as tandems, proximal, and dispersed duplicates in both modes. Differences arise between the modes in the following aspects -

'Overlap': Some genes, repeat across groups and this mode also produces an output file called 'Duplicates_relationships' that connects the different genes repeating across the duplicate groups

'Strict': Gene duplicates do not repeat across groups and this mode produces a fourth class of duplicates type called 'Mixed_duplicates' where the gene duplicates from different duplicate groups are merged together to retain their individual classification and also their relationship across the groups

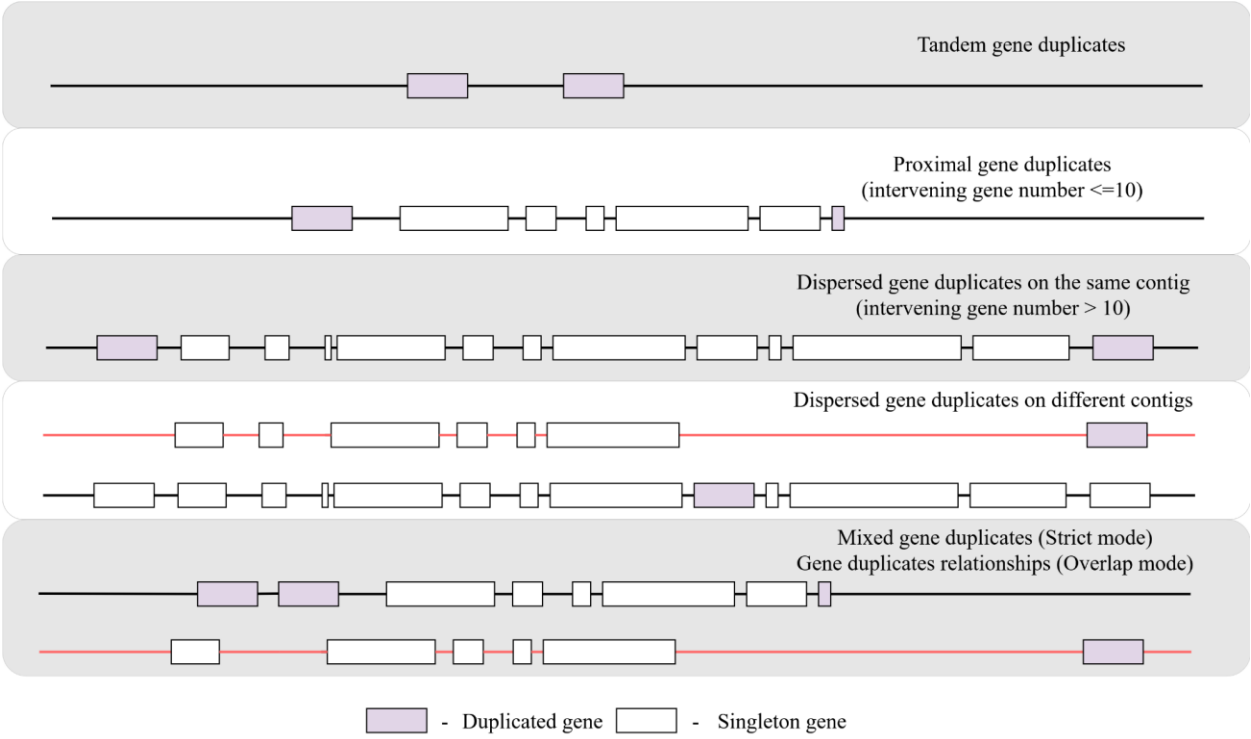

Figure 9: Different gene duplicate classifications output by DupyliCate. The red and black lines in the dispersed duplicates, mixed duplicates and gene duplicates relationships representations indicate different contigs/ scaffolds/ chromosomes.

The classification and clustering approach is detailed in Supplementary_Figure_4. As a clustering approach is used to obtain gene duplicate groups/arrays, for all the duplicate groups in the output, there is an internal confidence scoring scheme used to classify the gene duplicate group as low, moderate and high confidence group (Supplementary_Figure_5) that can help in interpreting the results accordingly. Along with the different duplicate group files, the singletons in a sample species are also provided per species.

In case the user has a set of desired genes in the reference whose copy number variation (CNV) needs to be assessed in the sample, a specific analysis flag option (--specific_genes) is available. Use of this flag produces two files. One is a comprehensive file providing information about the duplication and singleton status of the orthologs of the desired reference genes in the sample. Second is a copy number TSV file that provides the copy numbers of the specified reference genes across the different samples along with the orthologous group confidence score (OGCS) (OGCS scoring explained in Supplementary_Table_S9) to convey the confidence of the identified CNVs. This step also outputs a copy number XLSX file that has cells with no orthologs in the samples automatically colour coded in red, to allow for easy visual interpretation of presence-absence variation of the specified reference genes across the species analyzed.

Step 8: Synteny analysis: In the presence of a reference, synteny analysis is also carried out for small scale gene duplicate groups of tandem and proximal duplicates, to add a confidence layer of genomic positional context with respect to the reference. Also, if a reference is involved, detailed gene duplicate group nature details including gene group expansion and *de novo*

duplication are provided for small scale gene duplicate groups of tandem and proximal duplicates in the output. More information on the synteny analysis and gene duplicate group nature attributes have been described in Supplementary_Figure_6 and Supplementary_Figure_7 respectively. In the reference-free mode, ortholog detection, synteny analysis, gene group nature analysis steps are skipped.

Step 9: Ka/Ks computation: Optional downstream evolutionary analyses like Ka/Ks computation is performed for all gene duplicates on an individual gene basis using either Nei Gojobori (NG)[58] or Modified Yang Nielsen (MYN)[59] methods. These methods have been reimplemented from Ka_Ks_Calculator[10,60]. There is also in-built code facilitating amino-acid to codon-alignments thus removing the need for external tools in this step. While an exact implementation of the NG method is available, the MYN method implemented is an approximation of the MYN method implemented in Ka_Ks_Calculator (Supplementary_Figure_8 and Supplementary_Results_1). Supplementary_Figure_9 details the computational steps involved in the Ka/Ks analysis.

Step 10: Gene duplicates expression analysis and statistical analysis: If expression data is available for the sample(s), optional expression analysis of gene duplicates can also be performed by DupyliCate. This step gives out comprehensive information about the correlation among genes in duplicate groups, generates pairwise gene expression plots in a matrix figure and also helps determine divergent expression among duplicates as shown in Supplementary_Figure_10. This step also outputs a list of perceived (putative) pseudogenes that have a low average expression value across samples. Since spatial information of gene expression in the different tissues is not involved in this analysis, genes with low average expression value across the samples are designated as perceived pseudogenes. The method used in this expression analysis was adopted and modified from the method implemented by Qiao and co-workers[9].

Benchmarking

Arabidopsis thaliana Col-0 and Nd-1 accessions were used to benchmark DupyliCate v1.0, DupGen_finder v1.0.0[9], and doubletrouble v1.8.0[15]. DupyliCate was used in the overlap mode with *A. thaliana* Col-0 as the sample. DupyliCate additionally offers species-specific cutoff for segregating singletons and duplicates using BUSCO. Hence two runs of the tool were performed with *Arabidopsis thaliana* Col-0 dataset using BUSCO-based cutoff method and user-specified cut-off method. For the user-specified cutoff method results to be comparable with the other tools, a cutoff of 0 was specified for both normalized bit score and self similarity parameter flags. Since the standard classification scheme of doubletrouble was the closest to the classification scheme of DupyliCate, it was adopted for the doubletrouble run with *A. thaliana* Col-0 as the target. Transposed duplicates (TDs) are not classified in the standard scheme, hence, the doubletrouble run involved the use of only the target species *A. thaliana* Col-0. For DupGen_finder, *A. thaliana* Nd-1 was the outgroup and *A. thaliana* Col-0 as the target, as DupGen_finder does not offer classification schemes without TD classification, necessitating the use of an outgroup species. BLAST was used as the aligner for all three tools with a resource allocation of 15 threads during the benchmarking runs to avoid differences due to different sensitivity parameters in DIAMOND.

DupyliCate mainly assigns duplicates to tandems, proximals, and dispersed groups. Therefore, all the other duplicate group classifications apart from tandem and proximal duplicates in DupGen_finder and doubletrouble were processed with custom Python scripts to obtain a single miscellaneous group ('Misc.' in Table. 1) that is comparable to the dispersed duplicates group in DupyliCate. The benchmarking was therefore done mainly with respect to the number of genes classified as tandem, proximal, and dispersed duplicates by the three tools. Also, the run time of the three tools was compared. Five replicate runs were performed on the *A. thaliana* dataset using each of the three tools to estimate the runtime variability given the use of the same computational resources. Since local alignment is a step included within a run of DupyliCate, the local alignment times were also included in the runtime for the other two tools to standardize the comparison. The additional codes, and commands used for running doubletrouble, DupGen_finder and performing the benchmarking analysis can be found in the GitHub repository of DupyliCate (https://github.com/ShakNat/DupyliCate).

Run time analyses

The runtime of DupyliCate depends on the selected mode and parameter settings. The following approach was adopted to offer a comprehensive run time analysis when using the different modes and parameter settings. The run time for each run of the tool using *A. thaliana* datasets was recorded for the following run variations: 1) Overlap and strict mode; 2) Phylogeny and non-phylogeny-based ortholog detection in the presence of a reference; 3) Expression analysis of all identified gene duplicates; 4) Expression analysis of specified gene duplicates (XPB tandem duplication[17], BLA[18] and RALF tandem duplicates array[19] reported in *A. thaliana* Col-0 were chosen for demonstrating the specific expression analysis feature) 5) Ka/Ks analysis of all identified gene duplicates; and 6) Run with different aligners - BLAST/ DIAMOND/ MMseqs2.

Next, to analyze the tool run time increase with an increasing number of input datasets, a comprehensive run time analysis was conducted with the Brassicales plants used for the FLS case study. The different datasets were randomly grouped into groups of sizes 2, 4, 6, 8, 10, and 12, respectively. Run times were recorded using each group as an input for DupyliCate. To account for the variation in the file sizes of the datasets in each group, there were five random sets of the same group length generated, and the run time was recorded for every such replicate run.

Finally, to get an idea of the performance of DupyliCate in large-scale studies, the run time was also analyzed for the tool run on the 153 plant datasets used for the MYB case study that was done in the overlap mode using DIAMOND as the aligner, *A. thaliana* as the reference, and integrated phylogeny for ortholog detection.

Data Availability

The tool, the helper scripts, and test data sets can be accessed at: https://github.com/ShakNat/DupyliCate. Underlying data sets are publicly available at the NCBI,

Phytozome, and other databases as specified in Supplementary_Table_S1. Details of the supplementary data can be found in Supplementary_Information.

AI Statement

AI tools (ChatGPT and Claude) were used for optimizing, restructuring, and debugging the DupyliCate script.

Reference

1. Lallemand, T., Leduc, M., Landès, C., Rizzon, C. & Lerat, E. An Overview of Duplicated Gene Detection Methods: Why the Duplication Mechanism Has to Be Accounted for in Their Choice. *Genes* **11**, 1046 (2020).
2. Zhang, X. & Smith, D. R. An overview of online resources for intra-species detection of gene duplications. *Front. Genet.* **13**, (2022).
3. Durand, D. & Hoberman, R. Diagnosing duplications – can it be done? *Trends Genet.* **22**, 156–164 (2006).
4. Conant, G. C. & Wagner, A. GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res.* **30**, 3378–3386 (2002).
5. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* **13**, 2178–2189 (2003).
6. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
7. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
8. Djedatin, G., Monat, C., Engelen, S. & Sabot, F. *DuplicationDetector*, a light weight tool for duplication detection using NGS data. *Curr. Plant Biol.* **9–10**, 23–28 (2017).

9. Qiao, X. *et al.* Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **20**, 38 (2019).

10. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. *Genomics Proteomics Bioinformatics* **8**, 77–80 (2010).

11. Zhang, X., Hu, Y. & Smith, D. R. HSDFinder: A BLAST-Based Strategy for Identifying Highly Similar Duplicated Genes in Eukaryotic Genomes. *Front. Bioinforma.* **1**, (2021).

12. Zhang, X., Hu, Y., Smith, D. R., Cheng, Z. & Archibald, J. M. HSDSnake: a user-friendly SnakeMake pipeline for analysis of duplicate genes in eukaryotic genomes. *Bioinformatics* **41**, btaf325 (2025).

13. Sun, P. *et al.* WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* **15**, 1841–1851 (2022).

14. Kirilenko, B. M. *et al.* Integrating gene annotation with orthology inference at scale. *Science* **380**, eabn3107 (2023).

15. Almeida-Silva, F. & Van de Peer, Y. doubletrouble: an R/Bioconductor package for the identification, classification, and analysis of gene and genome duplications. *Bioinformatics* btaf043 (2025) doi:10.1093/bioinformatics/btaf043.

16. Hanke, D. M. & Dagan, T. SegMantX: a novel tool for detecting DNA duplications uncovers prevalent duplications in plasmids. Preprint at https://doi.org/10.1101/2025.03.14.643293 (2025).

17. Masuda, H. P. *et al.* Evidence for sub-functionalization of tandemly duplicated *XPB* nucleotide excision repair genes in *Arabidopsis thaliana*. *Gene* **754**, 144818 (2020).

18. Ugalde, J. M. & Straube, H. New genes on the block: Neofunctionalization of tandem duplicate genes with putative new functions in Arabidopsis. *Plant Physiol.* **192**, 2574–2576 (2023).

19. Cao, J. & Shi, F. Evolution of the RALF Gene Family in Plants: Gene Duplication and Selection Patterns. *Evol. Bioinforma. Online* **8**, 271–92 (2012).

20. Pucker, B. *et al.* A De Novo Genome Sequence Assembly of the Arabidopsis thaliana Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny. *PLOS ONE* **11**, e0164321 (2016).

21. Pucker, B. *et al.* A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. *PLOS ONE* **14**, e0216233 (2019).

22. Xu, Z. *et al.* Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. *BMC Biol.* **18**, 63 (2020).

23. Sielemann, K. *et al.* Genomic characterization of a nematode tolerance locus in sugar beet. *BMC Genomics* **24**, 748 (2023).

24. Hakim, S. E. *et al.* Phylogenomics and metabolic engineering reveal a conserved gene cluster in Solanaceae plants for withanolide biosynthesis. *Nat. Commun.* **16**, 6367 (2025).

25. Barker, M. S., Vogel, H. & Schranz, M. E. Paleopolyploidy in the Brassicales: Analyses of the Cleome Transcriptome Elucidate the History of Genome Duplications in Arabidopsis and Other Brassicales. *Genome Biol. Evol.* **1**, 391–399 (2009).

26. Charon, C., Bruggeman, Q., Thareau, V. & Henry, Y. Gene duplication within the Green Lineage: the case of TEL genes. *J. Exp. Bot.* **63**, 5061–5077 (2012).

27. Gao, B. *et al.* Evolution by duplication: paleopolyploidy events in plants reconstructed by deciphering the evolutionary history of VOZ transcription factors. *BMC Plant Biol.* **18**, 256 (2018).

28. Blanc, G. & Wolfe, K. H. Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution. *Plant Cell* **16**, 1679–1691 (2004).

29. Vukašinović, N. *et al.* Dissecting a Hidden Gene Duplication: The Arabidopsis thaliana SEC10 Locus. *PLOS ONE* **9**, e94077 (2014).

30. Woollard, A. Gene duplications and genetic redundancy in C. elegans. in *WormBook: The Online Review of C. elegans Biology [Internet]* (WormBook, 2005).

31. Delaux, P.-M., Séjalon-Delmas, N., Bécard, G. & Ané, J.-M. Evolution of the plant–microbe symbiotic 'toolkit'. *Trends Plant Sci.* **18**, 298–304 (2013).

32. Emms, D. M. & Kelly, S. SHOOT: phylogenetic gene search and ortholog inference. *Genome Biol.* **23**, 85 (2022).

33. Schilbert, H. M. *et al.* Characterization of the Brassica napus Flavonol Synthase Gene Family Reveals Bifunctional Flavonol Synthases. *Front. Plant Sci.* **12**, (2021).

34. Schilbert, H. M. *et al.* Characterization of the Brassica napus Flavonol Synthase Gene Family Reveals Bifunctional Flavonol Synthases. *Front. Plant Sci.* **12**, (2021).

35. Schilbert, H. M. & Glover, B. J. Analysis of flavonol regulator evolution in the Brassicaceae reveals MYB12, MYB111 and MYB21 duplications and MYB11 and MYB24 gene loss. *BMC Genomics* **23**, 604 (2022).

36. Davies, K. M. *et al.* The Evolution of Flavonoid Biosynthesis: A Bryophyte Perspective. *Front. Plant Sci.* **11**, (2020).

37. Davies, K. M. *et al.* The evolution of flavonoid biosynthesis. *Philos. Trans. R. Soc. B Biol. Sci.* **379**, 20230361 (2024).

38. Yonekura-Sakakibara, K., Higashi, Y. & Nakabayashi, R. The Origin and Evolution of Plant Flavonoid Metabolism. *Front. Plant Sci.* **10**, 943 (2019).

39. Mao, L. *et al.* RiceRelativesGD: a genomic database of rice relatives for rice research. *Database* **2019**, baz110 (2019).

40. Karberg, K. A., Olsen, G. J. & Davis, J. J. Similarity of genes horizontally acquired by Escherichia coli and Salmonella enterica is evidence of a supraspecies pangenome. *Proc. Natl. Acad. Sci.* **108**, 20154–20159 (2011).

41. Cliften, P. F., Fulton, R. S., Wilson, R. K. & Johnston, M. After the Duplication: Gene Loss and Adaptation in Saccharomyces Genomes. *Genetics* **172**, 863–872 (2006).

42. Ma, F., Lau, C. Y. & Zheng, C. Young duplicate genes show developmental stage- and cell type-specific expression and function in Caenorhabditis elegans. *Cell Genomics* **4**, 100467 (2024).

43. Owens, D. K. *et al.* Functional Analysis of a Predicted Flavonol Synthase Gene Family in Arabidopsis. *Plant Physiol.* **147**, 1046–1061 (2008).

44. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

45. Brown, J. W., Walker, J. F. & Smith, S. A. Phyx: phylogenetic tools for unix. *Bioinformatics* **33**, 1886–1888 (2017).

46. Wong, T. K. F. *et al.* IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary Models. https://ecoevorxiv.org/repository/view/8916/ (2025).

47. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

48. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

49. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).

50. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. in *Gene Prediction: Methods and Protocols* (ed. Kollmar, M.) 161–177 (Springer, New York, NY, 2019). doi:10.1007/978-1-4939-9173-0_9.

51. Dainat, J. *et al.* NBISweden/AGAT: AGAT v1.5.1. Zenodo https://doi.org/10.5281/zenodo.16317950 (2025).

52. Tegenfeldt, F. *et al.* OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res.* **53**, D516–D522 (2025).

53. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

54. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

55. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

56. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).

57. Sukumaran, J. & Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010).

58. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).

59. Wang, D.-P., Wan, H.-L., Zhang, S. & Yu, J. γ-MYN: a new algorithm for estimating Ka and Ks with consideration of variable substitution rates. *Biol. Direct* **4**, 20 (2009).

60. Zhang, Z. KaKs_Calculator 3.0: Calculating Selective Pressure on Coding and Non-Coding Sequences. *Genomics Proteomics Bioinformatics* **20**, 536–540 (2022).

Acknowledgements

The authors thank the Plant Biotechnology and Bioinformatics group (University of Bonn) for valuable inputs and inspiring discussions. SN was supported by the doctoral scholarship of the German Academic Exchange Service (DAAD). This work was supported by the de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) and ELIXIR-DE (Forschungszentrum Jülich and W-de.NBI-001, W-de.NBI-004, W-de.NBI-008, W-de.NBI-010, W-de.NBI-013, W-de.NBI-014, W-de.NBI-016, W-de.NBI-022). Open access funding enabled by Project DEAL.

Funding

Open access funding enabled by Project DEAL. There was no specific funding for this research.

Author information

Authors and Affiliations

Plant Biotechnology and Bioinformatics, Institute for Cellular and Molecular Botany, University of Bonn, Kirschallee 1, 53115, Bonn, Germany
Shakunthala Natarajan & Boas Pucker

Contributions

S.N. and B.P. designed the study. S.N. wrote the software and conducted the bioinformatic analyses. S.N. and B.P. interpreted the results and wrote the manuscript. All authors read the final version of the manuscript and agreed to its submission.

Corresponding author

Correspondence to pucker@uni-bonn.de

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.