Pak. J. Bot., 59(1): January 2027.

ARTICLE | DOI: http://dx.doi.org/10.30848/PJB2027-1(15)

COMPARATIVE GENOMICS OF SEVEN NUCLEAR ASSEMBLIES SPANNING FOUR BORAGINALES FAMILIES REVEALS A PALEOPOLYPLOIDY LANDSCAPE AND TRANSPOSON-ASSOCIATED ASSEMBLY-SIZE VARIATION

WENCAI JIE[1,2#], YONGHUI LIAO[1,2#], SHOUCHENG HUANG[1,2,3#], XIAOHUI LAI[1,2,4#], YUXIN ZHANG[1,2], XIAORAN LV[1,2], CHANGYI WANG[1,2], TONGMING YIN[2], MINKAI YANG[1,2], ZHONGLING WEN[1,2], HONGWEI HAN[1,2,5], JINLIANG QI[1,2], GUIHUA LU[1,2,6]* AND YONGHUA YANG[1,2]*

[1]Institute for Plant Molecular Biology, State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing 210023, China
[2]Co-Innovation Center for Sustainable Forestry in Southern China, State Key Laboratory for Tree Genetics and Breeding, Nanjing Forestry University, Nanjing 210037, China
[3]College of Agriculture, Anhui Science and Technology University, Fengyang 233100, China
[4] College of Biological Sciences and Technology, Yili Normal University, Yining 835000, China
[5] School of Life Sciences and Chemical Engineering, Jiangsu Second Normal University, Nanjing 210013, China
[6]School of Life Sciences, Huaiyin Normal University, Huaian 223300, China
[#]Contributed equally

*Corresponding author's email: ghlu@hytc.edu.cn; yangyh@nju.edu.cn

Abstract

An integrated comparison of transposable-element landscapes, paleopolyploidy signals and macro-synteny across Boraginales families remains limited. We analysed seven nuclear assemblies from four families spanning an 11.5-fold range in raw analysed assembly size (158.3–1,822.6 Mb). TE content ranged from 13.31% to 76.81% of non-N sequence and was strongly associated with $\log_{10}$ assembly size ($R^2 = 0.937$, $p<0.001$; $n = 7$). Intact-LTR age profiles differed among lineages, while their ecological and causal implications remain untested. A uniform paranome and self-synteny sensitivity analysis identified anchor-supported, window-stable candidate components in the sampled *L. erythrorhizon* and *P. sempervirens* genomes; components in the other five genomes remained descriptive because at least one candidate-classification condition failed. These within-genome results do not establish a shared ancestral WGD in Boraginaceae or precise event dates. Gene-based macro-synteny remained detectable across the four families despite their broad assembly-size range.

Key words: Boraginales; Comparative genomics; Macro-synteny; Paleopolyploidy; Transposable elements; Whole-genome duplication

Introduction

Boraginales comprise approximately 2,700 species and have traditionally been treated as 11 families (Luebert *et al*., 2016), although recent phylogenomic work has proposed a revised nine-family framework (Vasile *et al*., 2025). Representative taxa include the medicinal *Lithospermum erythrorhizon*, the invasive *Echium plantagineum*, and tropical woody genera such as Cordia and Ehretia. Although publicly available Boraginales assemblies have increased, the subset suitable for a uniform comparison remains limited by annotation availability and assembly usability. We therefore analysed seven nuclear assemblies from four families, spanning 158.3–1,822.6 Mb in raw analysed assembly size (11.5-fold), under a common framework for TE annotation, paranome K_s analysis and macro-synteny.

Angiosperm genome sizes vary across more than three orders of magnitude, reflecting polyploidy, TE proliferation and genome contraction (Lee & Kim, 2014; Michael, 2014). LTR retrotransposon abundance contributes substantially to genome-size variation across angiosperms (Wang *et al*., 2021), while repeated ancient duplications are widespread in asterids (Landis *et al*., 2018; Zhang *et al*., 2020). Cross-family sampling is needed to distinguish observations shared among sampled lineages from signals recovered in only one genome. Previous studies placed ancient-duplication signals at several levels within Boraginales, but their taxon sampling, K_s construction, event-placement methods and rate calibrations differ (Auber *et al*., 2020; Tang *et al*., 2020; Zhang *et al*., 2020). Their numerical estimates are therefore treated here as methodological context rather than as directly interchangeable event ages. Subsequent genome releases expanded the available taxonomic scope to *L. erythrorhizon* (Auber *et al*., 2020), *Cordia subcordata* (Chen *et al*., 2023), *E. macrophylla* (Cheng *et al*., 2024), and *P. recurvata* (Northing *et al*., 2025).

Among the selected assemblies, *E. plantagineum*, *Pentaglottis sempervirens* and *Nemophila menziesii* lacked public gene models suitable for the comparative workflow, and we therefore generated de novo annotations for these three genomes. We integrated them with four published annotations to address two questions: (Q1) How is TE content associated with the 11.5-fold variation in analysed assembly size? (Q2) Which within-genome paleoduplication components withstand a uniform sensitivity screen, and how much gene-based macro-synteny is retained across the sampled families?

Received: 08-05-2026 | Revised: 01-08-2026 | Accepted: 08-08-2026 | Online: 30-08-2026

Material and Methods

Genome data sources: Seven Boraginales species from four families were included, with *Solanum lycopersicum* as outgroup (Table 1). "Assembly size" denotes raw length of the FASTA sequence set used in the comparative analyses, including N bases. For *P. recurvata*, this is the primary chromosome/contig fraction (158.3 Mb) rather than the full deposited record. Three genomes lacking suitable public gene models were annotated de novo: *E. plantagineum* (GCA_003412495.2), *P. sempervirens* (GCA_964197855.1), and *N. menziesii* (GCA_029959085.1; Shaffer *et al*., 2022). The remaining species retained published assemblies and annotations: *L. erythrorhizon* (Purdue assembly GCA_019722365.1; Auber *et al*., 2020), *C. subcordata* (Chen *et al*., 2023), *E. macrophylla* (the 33-sequence haplotype-a analysis set; Cheng *et al*., 2024), and *P. recurvata* (Northing *et al*., 2025). Family circumscription follows Luebert *et al*., (2016) and the updated framework of Vasile *et al*., (2025), in which *Ehretia* is placed in Ehretiaceae.

Gene annotation: A unified multi-evidence pipeline was applied to *E. plantagineum*, *P. sempervirens* and *N. menziesii*. Homology-based prediction used GeMoMa v1.9 (Keilwagen *et al*., 2018; 2019) with proteins from representative asterid models; transcript evidence was assembled with Trinity and StringTie (Grabherr *et al*., 2011; Pertea *et al*., 2015) from public RNA-seq data (10 samples for *E. plantagineum*, seven for *N. menziesii* and one for *P. sempervirens*), and ab initio prediction used AUGUSTUS (Stanke *et al*., 2006). EVidenceModeler and two PASA update rounds integrated the evidence, and functional annotation combined InterProScan v5 with DIAMOND searches against major protein databases (Haas *et al*., 2008; Jones *et al*., 2014; Buchfink *et al*., 2021). Completeness was assessed with BUSCO v6 under eudicotyledons_odb12 (2,805 loci; Manni *et al*., 2021; Tegenfeldt *et al*., 2025) in genome and protein modes. For the three newly annotated species, complete BUSCO values were 98.0/92.9% for *E. plantagineum*, 94.5/91.3% for *P. sempervirens* and 93.8/82.3% for *N. menziesii* (genome/protein mode). The Purdue *L. erythrorhizon* inputs scored 94.4/77.1%. Genome mode and protein mode assess non-equivalent inputs, and the retained records do not allow their differences to be attributed solely to assembly completeness or solely to gene-model recovery.

Transposable element annotation: The retained TE-annotation outputs for the seven analysed assemblies identify EDTA v2.3 (Ou *et al*., 2019; 2024). Their summaries were parsed under one reporting rule. EDTA was selected because its plant-focused workflow combines structure-based and homology-based discovery and applies one classification scheme to intact and fragmented elements. The reported TE percentage is the exact sum of LTR/Copia, LTR/Gypsy, other LTR, DNA transposon, LINE, SINE and unclassified base pairs divided by non-N FASTA length; rDNA and snRNA were excluded. Raw FASTA length, including N, was recorded separately as assembly size. For each retained intact element, the actual analysis defined uncorrected paired-LTR divergence as $p = 1 - \text{LTR identity}$ and calculated an identity-derived insertion-age proxy as $T = p/(2r)$, using $r = 1.3 \times 10^{-8}$ substitutions site^{-1} year^{-1}. The rate follows Ma and Bennetzen (2004), whereas the distance used here is not Kimura two-parameter corrected. The TE–assembly-size relationship was tested by ordinary least squares with TE percentage as the response and $\log_{10}$ raw assembly size in Mb as the predictor; the P value is two-sided, and seven leave-one-out refits assessed sensitivity to individual species.

Whole-genome duplication analysis: Paranome duplication signals were analysed with the wgd workflow (Zwaenepoel & Van de Peer, 2019): wgd dmd constructed paralog families and wgd ksd estimated synonymous-substitution values. For the mixture analysis, the filtering order matched wgd v2.0.38 and one CDS set matched to its gene-coordinate file was used per species. *L. erythrorhizon* was restricted to the Purdue assembly and its 26,117-gene set, for which DMD and KSD were rerun. Complete KSD records with alignment length ≥300 were retained when $0 < K_s \leq 5$, averaged by paralog family and node, and transformed as $\ln(K_s)$. Full-covariance Gaussian mixture models with K = 1–4 were fitted in scikit-learn (Pedregosa *et al*., 2011) with 200 initialisations, a maximum of 200 iterations and fixed species-level seeds. All selected fits converged. BIC was primary and AIC secondary for selecting K; component location is reported as $\exp(\mu)$, where μ is the fitted mean on the ln (K_s) scale. No alternative clustering family was tested, so fitted components are treated as descriptive unless they pass the independent sensitivity screen below.

Table 1. Overview of seven Boraginales species included in this study.

Species	Family	Assembly size (Mb)[a]	TE content (%)[a]	Screened component[b]	Gene count[c]	Source[d]
L. erythrorhizon	Boraginaceae	367	43.93	0.392	26,117	Downloaded
E. plantagineum	Boraginaceae	349	30.40	Not assigned	33,946	This study
P. sempervirens	Boraginaceae	851	63.57	0.324	33,814	This study
P. recurvata	Boraginaceae	158	13.31	Not assigned	30,655	Downloaded
C. subcordata	Cordiaceae	472	48.49	Not assigned	26,616	Downloaded
E. macrophylla	Ehretiaceae	1,823	76.81	Not assigned	29,660	Downloaded
N. menziesii	Hydrophyllaceae	1,522	65.93	Not assigned	27,053	This study

[a] Assembly size is raw analysed FASTA length, including N, rounded to the nearest megabase. TE content is the exact sum of LTR/Copia, LTR/Gypsy, other LTR, DNA transposon, LINE, SINE and unclassified base pairs divided by non-N FASTA length; rDNA and snRNA are excluded. The *E. macrophylla* value refers to the 33-sequence haplotype-a analysis set. [b] A component is listed only when it met the four within-genome candidate conditions defined in Methods; "Not assigned" means that no screened component passed all conditions and is not a claim that WGD is absent. [c] Gene counts are those in the comparative dataset. [d] "This study" denotes annotations produced de novo; "Downloaded" denotes gene-model sets from the original source

Numerical reproducibility was checked in two complete fixed-seed runs, and the screened component in each species was assessed across seven predefined alignment/window settings and against independently rerun self-synteny anchors. A component was classified as a within-genome paleoduplication candidate only if four conditions were met: (i) at least 70% and at least 200 self-synteny anchor pairs mapped to the retained node-averaged K_s dataset; (ii) component-location change was ≤15% across the seven fits; (iii) exp(μ) fell within the intersection of the main anchor-node K_s peak full-width-at-half-maximum intervals obtained with SciPy's gaussian_kde implementation at 0.8×, 1.0× and 1.2× Scott bandwidth, with peak-centre shifts ≤15% (Scott, 1992; Virtanen *et al*., 2020); and (iv) excluding genes with ≥50% merged-CDS overlap with merged EDTA intervals did not change the interpretation. For rate sensitivity, screened components were converted algebraically using $T = K_s/(2r)$. The value $r = 6.5 \times 10^{-9}$ substitutions site^{-1} year^{-1} is the rounded mean of the grass Adh estimates reported by Gaut *et al*., (1996) and was also adopted by Ma and Bennetzen (2004); 5.2 and 7.8×10^{-9} are ±20% scenarios around that value, and 8.84×10^{-9} follows Tang *et al*., (2020). These scenarios show dependence on the assumed rate; they neither estimate lineage-specific rates nor establish event dates.

Synteny analysis: Twenty-one pairwise and seven self-synteny JCVI MCscan analyses were used (Tang *et al*., 2008; 2024). The cross-species comparisons used LAST-based gene matching (Kiełbasa *et al*., 2011), cscore = 0.7 and a minimum block size of four anchors; within-genome self-synteny analyses used cscore = 0.7 and a minimum of five anchors. The retained run records used nucleotide CDS FASTA files and BED-formatted gene coordinates rather than whole-genome FASTA sequences, so no separate whole-genome soft- or hard-masking step was applied. Complete command records are unavailable for some cross-species comparisons, and a common TE-related gene prefilter cannot be confirmed across them. These outputs are therefore interpreted descriptively as gene-based collinearity and were not used to classify the candidate components.

Phylogenomic inference: OrthoFinder v3.1.3 (Emms & Kelly, 2019) clustered proteins from the eight taxa (seven Boraginales plus S. lycopersicum). The 839 single-copy orthologues were aligned with MAFFT (Katoh & Standley, 2013), trimmed with ClipKIT in gappy mode (Steenwyk *et al*., 2020), and concatenated into a 540,634-amino-acid supermatrix. Maximum-likelihood inference was run in IQ-TREE 3 under LG+G4 (Wong *et al*., 2026) with 1,000 ultrafast bootstrap replicates (Hoang *et al*., 2018).

Results

TE content is associated with the 11.5-fold assembly-size variation: Across the seven assemblies, TE content calculated against non-N sequence ranged from 13.31% in *P. recurvata* (158.3 Mb) to 76.81% in *E. macrophylla* (1,822.6 Mb); *N. menziesii* and *P. sempervirens* contained 65.93% and 63.57%, respectively (Fig. 1a,b). TE content was strongly associated with $\log_{10}$ raw assembly size (R^2 = 0.937; two-sided p=0.000343; n = 7). The association remained positive in every leave-one-out refit (R^2 = 0.893–0.954; all p<0.005), although the observations are phylogenetically non-independent. TE composition also differed sharply: LTR retrotransposons accounted for 69.91% of non-N sequence in *E. macrophylla* and 56.32% in *P. sempervirens*, *N. menziesii* contained a large LINE fraction (20.89%), and the unclassified fraction was highest in *L. erythrorhizon* (22.61%). These categories describe annotated sequence composition and should not be read as rates of gain or loss.

Median identity-derived insertion-age estimates for retained intact LTR elements were 0.04 Ma in *E. plantagineum*, 0.21 Ma in *P. sempervirens*, 0.37 Ma in *N. menziesii*, 0.88 Ma in *P. recurvata*, 0.98 Ma in *E. macrophylla*, 1.35 Ma in *C. subcordata* and 2.39 Ma in *L. erythrorhizon* (Fig. 1c). The contrasting distributions indicate lineage-specific concentrations of insertion and retention ages among intact elements; they do not measure complete TE turnover or current activity. *P. recurvata* combined the smallest analysed assembly with the lowest TE fraction. Its arid annual life history is relevant biological context, but the present analysis did not test ecological adaptation or the mechanism of genome compaction.

Paleopolyploidy landscape across the seven sampled Boraginales genomes: The species tree inferred from 839 single-copy orthologues (Fig. 2a; 540,634-amino-acid supermatrix; 100% ultrafast bootstrap support at every internal node) resolves the four-family relationships in a topology compatible with the sampled family-level relationships recovered by Vasile *et al*., (2025). It provides phylogenetic context for the within-genome component summary in Fig. 2b; the components are not mapped to ancestral branches. *N. menziesii* showed the longest terminal branch (0.250 substitutions per site).

Panel a shows the maximum-likelihood species tree inferred from 839 single-copy orthologues, with *S. lycopersicum* as outgroup. Panel b shows the screened exp(μ) component for each genome. Filled points are the *L. erythrorhizon* and *P. sempervirens* components that m*et al*l four candidate conditions; open points are descriptive components for which at least one condition failed. Point colours denote families. The panel is not an absolute-age WGD timeline and does not map a shared event to an ancestral branch.

Among the four sampled Boraginaceae genomes, *L. erythrorhizon* and *P. sempervirens* contained candidate components at K_s = 0.392 and 0.324, respectively (Figs. 2b and 3; Table 1). Their maximum component-location changes across the predefined fits were 3.9% and 2.7%, and 74.5% (2,239 pairs) and 87.4% (5,068 pairs) of their self-synteny anchor pairs were retained in the filtered node-averaged K_s datasets. The *E. plantagineum* (K_s = 0.659) and *P. recurvata* (K_s = 0.668) components remained descriptive because their maximum changes were 27.9% and 51.2% and neither fell within the bandwidth-stable anchor-peak interval (Fig. 3). Self-synteny provides complementary within-genome evidence but does not independently date a component or establish shared ancestry. Tang *et al*., (2020) proposed Boraginaceae signals near $K_s \approx 0.417$ and 0.939 using mixed genome/transcriptome data, whereas Auber *et al*., (2020) observed a syntenic-pair peak near 0.45 in *L. erythrorhizon*; these are qualitative context rather than method-equivalent estimates.

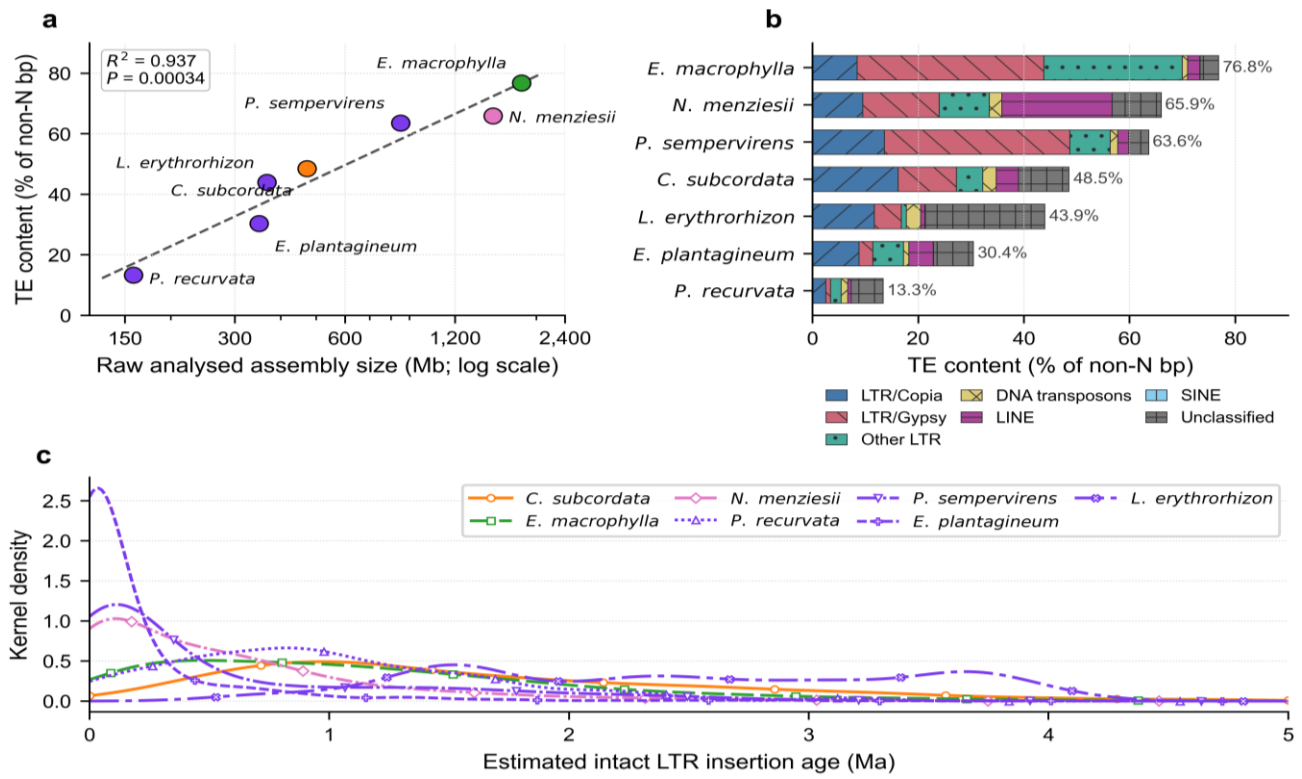

Fig. 1. Assembly size, TE composition and intact-LTR insertion-age distributions across seven Boraginales genomes.
(a) TE content, calculated as the exact seven-class TE sum divided by non-N FASTA base pairs, plotted against raw analysed assembly size. The x-axis is logarithmic; the dashed line is the ordinary least-squares fit (n = 7; R^2 = 0.937; two-sided P = 0.000343). Point colours denote families. (b) Contributions of LTR/Copia, LTR/Gypsy, other LTR, DNA transposons, LINE, SINE and unclassified elements to the same denominator; rDNA and snRNA are excluded. (c) Kernel-density estimates from 0 to 5 Ma for identity-derived insertion-age proxies of retained intact LTR elements, calculated with p = 1 − paired-LTR identity and T = p/(2r). Local density maxima indicate concentrations of inferred insertion and retention ages, not direct measurements of current activity or complete TE turnover. Species are distinguished redundantly by colour, line style and point symbol.

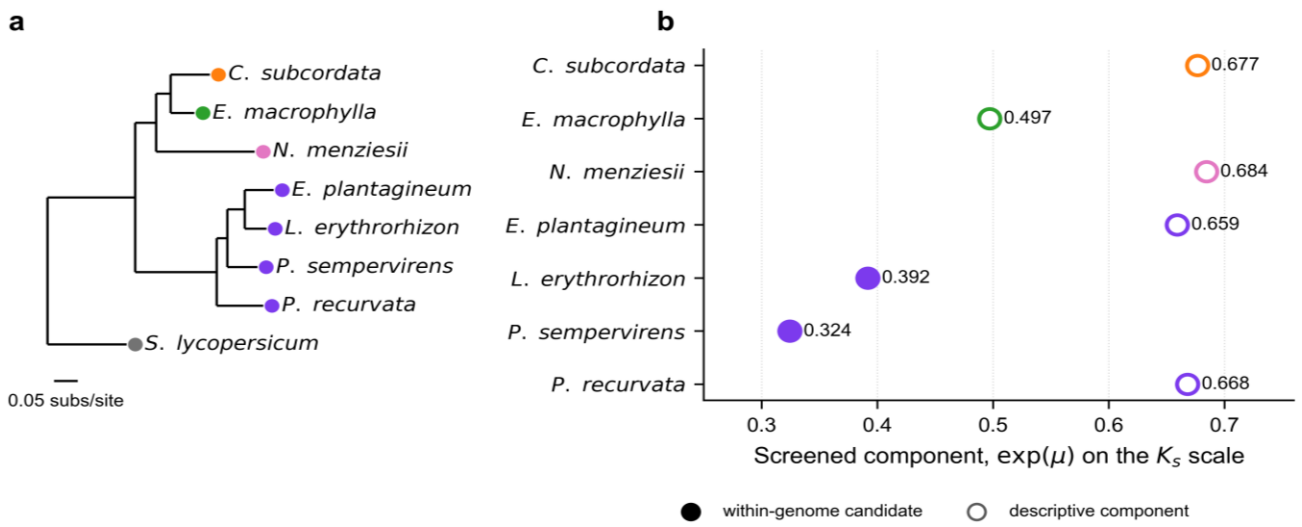

Fig. 2. Species tree and screened within-genome paleoduplication components for seven Boraginales genomes.

The *C. subcordata* component was centred at K_s = 0.677 and remained descriptive because its location changed by 21.4% across fitting settings. The *E. macrophylla* component was centred at K_s = 0.497; its location changed by 14.8% and coincided with the stable anchor-K_s interval, but only 66.7% (5,454 of 8,182) of self-synteny anchor pairs remained after the complete-record, alignment-length and K_s filters, below the 70% criterion. It is therefore also descriptive. Algebraic conversion of K_s = 0.497 gives 38.2 Ma at r = 6.5 × 10^{-9} and 28.1 Ma at r = 8.84 × 10^{-9} substitutions site^{-1} year^{-1}; these values illustrate rate sensitivity and are not event-age estimates. Tang *et al*., (2020) reported $K_s \approx 0.675$ and 38.17 Ma for transcriptome-derived *E. microphylla* under a different workflow and rate. Species, input construction, component definition and calibration preclude a direct biological interpretation of the numerical contrast.

In the single sampled *N. menziesii* genome, the screened component was centred at $K_s = 0.684$, changed by 60.3% across fitting conditions and lay below the stable self-synteny anchor-peak interval of $K_s = 0.813–1.767$. We therefore assigned it neither to a younger nor to an ancestral duplication event. Overall, only two of the seven screened components met all four within-genome candidate conditions; the remaining five are reported descriptively rather than converted into lineage-level WGD claims.

Panel a compares each screened exp(μ) value with the bandwidth-stable full-width-at-half-maximum interval of the self-synteny anchor-node K_s peak. Panel b shows the maximum relative component-location change across seven predefined alignment/window fits; the dashed line marks the 15% condition. Panel c shows the proportion of self-synteny anchor pairs mapped after KSD filtering; the dashed line marks the 70% condition. Filled symbols/bars denote the two components that met all four candidate conditions; open symbols and grey bars denote descriptive components. The fourth condition, positional sensitivity to removal of genes with high merged-CDS/EDTA overlap, did not change any classification.

Cross-family macro-synteny conservation: Pairwise JCVI MCscan analyses recovered gene-based collinear blocks across the four sampled families (Fig. 4; representative ribbons in Fig. 5). The largest off-diagonal matrix count was *E. plantagineum* × *P. sempervirens* (28,762 anchors). Cross-family comparisons also retained substantial counts, including *N. menziesii* × *E. plantagineum* (14,545) and *E. macrophylla* × *E. plantagineum* (20,671). Block-level structure remained evident in the ribbon plots despite the broad range in analysed sequence-set size. Because complete command records are unavailable for some cross-species comparisons, these counts are used as a descriptive macro-synteny comparison rather than as a test of TE-filtered anchor recovery.

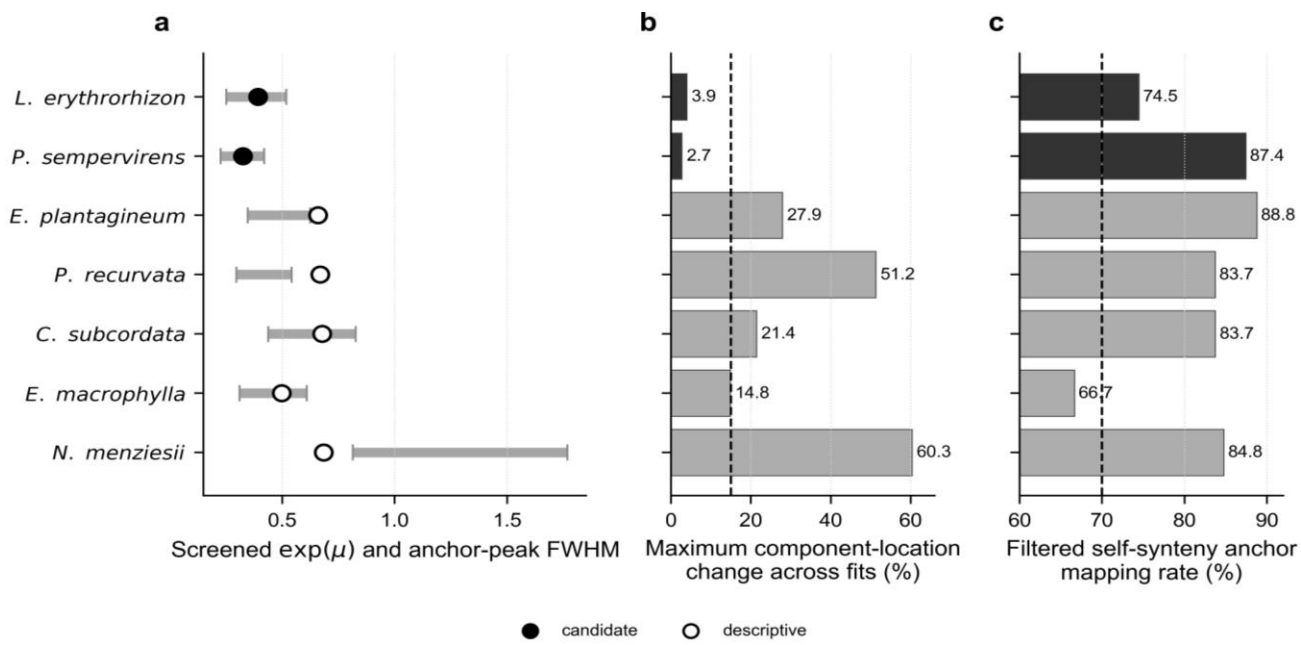

Fig. 3. Sensitivity diagnostics for the screened paranome K_s components.

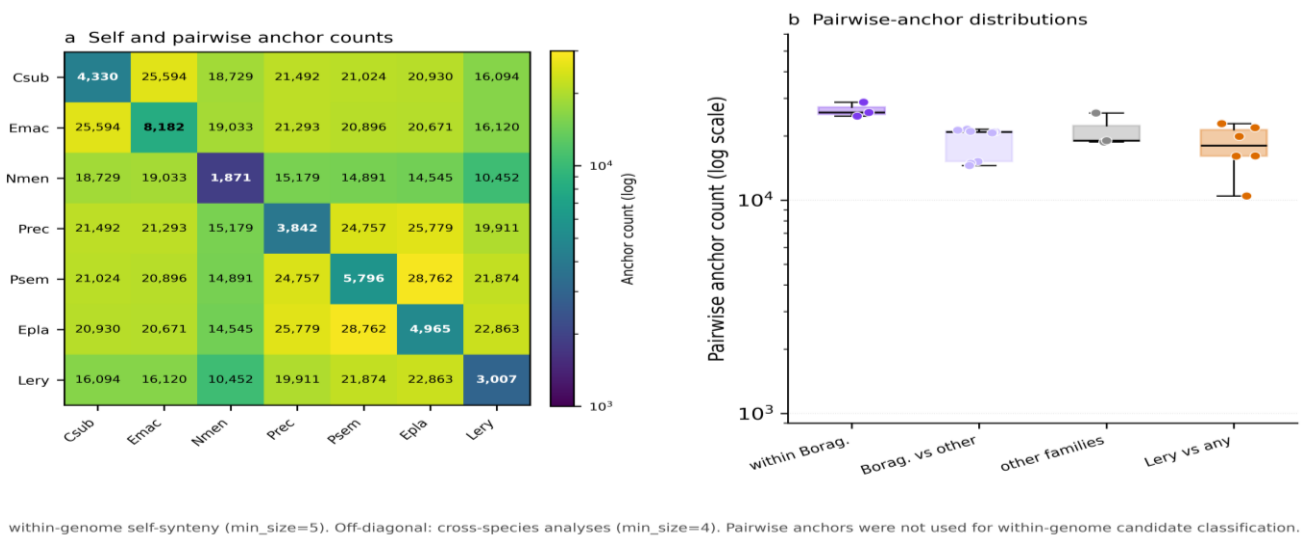

Fig. 4. Macro-synteny anchor recovery across seven Boraginales genomes.

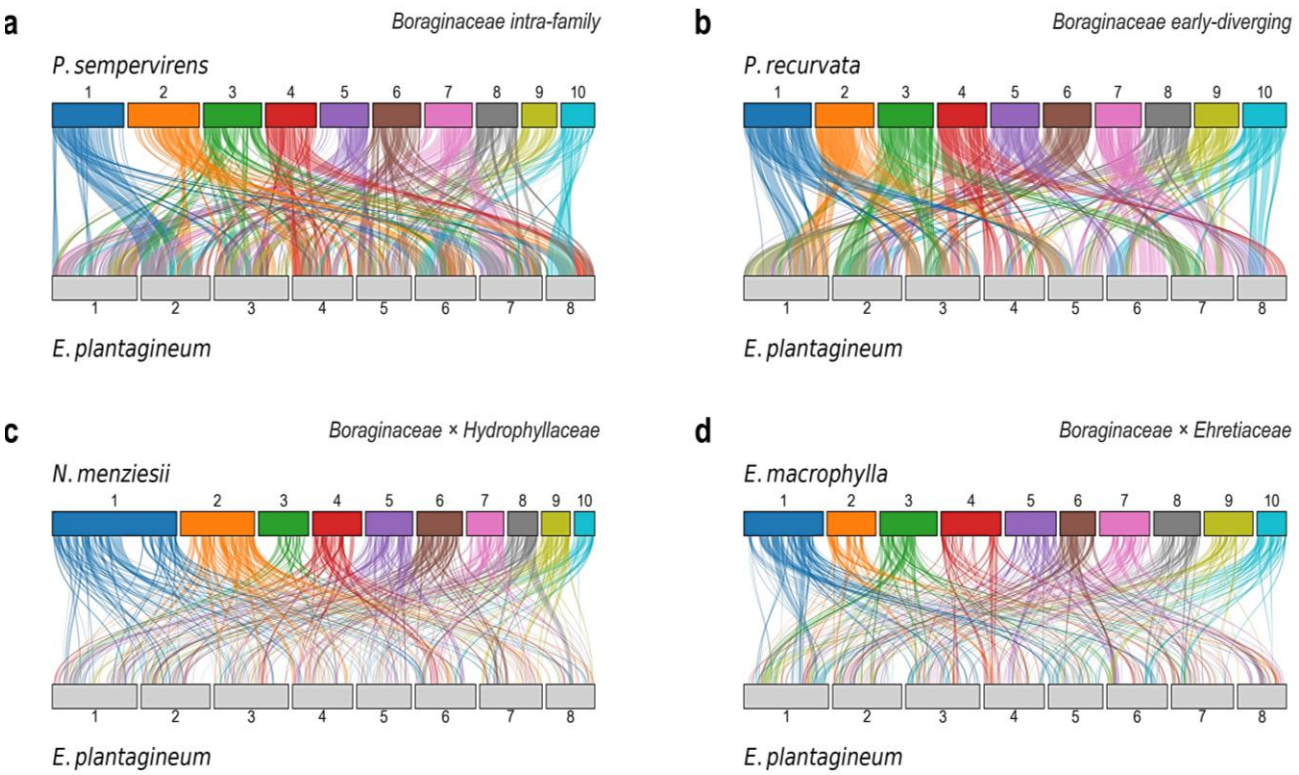

Fig. 5. Representative macro-synteny ribbons using *E. plantagineum* as the common reference.

Panel a shows the JCVI MCscan self- and pairwise-anchor matrix. Diagonal cells report within-genome self-synteny analyses (cscore = 0.7; minimum block size of five anchors), whereas off-diagonal cells report cross-species comparisons (cscore = 0.7; minimum block size of four anchors). Panel b summarizes pairwise anchor-count distributions by family context on a logarithmic scale. The off-diagonal comparisons are descriptive and were not used to classify the candidate components in Fig. 3.

Karyotype ribbons show JCVI MCscan blocks between each focal species and *E. plantagineum* LG1–LG8. Only blocks with ≥10 anchor genes confined to one top-track chromosome are drawn and coloured by the top-track chromosome. Panels show (a) *P. sempervirens* × *E. plantagineum*, (b) *P. recurvata* × *E. plantagineum*, (c) *N. menziesii* × *E. plantagineum*, and (d) *E. macrophylla* × *E. plantagineum*. Drawn-anchor counts after block and top-track filtering differ from the non-redundant pairwise counts in Fig. 4. The Purdue *L. erythrorhizon* input is a fragmented, non-chromosome-scale contig assembly without chromosome-level pseudomolecules and was therefore omitted from the ribbon panels; its self-synteny result remains represented in Fig. 4.

Discussion

TE content is strongly associated with assembly-size variation: TE content was strongly associated with raw assembly size in this seven-genome comparison. The association was insensitive to omission of any one species, but its cross-sectional and phylogenetically non-independent basis leaves the causal contribution of TE accumulation unresolved. LTR sequence predominated in the two largest analysed assemblies, consistent with substantial retention in their present sequence content. Distinguishing insertion, removal and host-silencing histories will require lineage-aware analyses and additional genomes.

The compact *P. recurvata* assembly and the TE-rich *E. macrophylla* and *N. menziesii* assemblies occupy opposite ends of the comparison. Habitat and life history were not tested as explanatory variables, leaving adaptive genome compaction in *P. recurvata* untested. Intact-LTR age profiles summarize elements retained and recognizable in each assembly. They identify concentrations of relatively young inferred ages without quantifying complete turnover, separating activity from differential removal, or demonstrating an ecological consequence.

Paleopolyploidy landscape in cross-family context: Earlier transcriptome and single-genome studies inferred ancient-duplication signals in Boraginaceae. In the present comparison, only *L. erythrorhizon* and *P. sempervirens* met the four-condition within-genome candidate rule. The *E. plantagineum* and *P. recurvata* components failed the same rule, leaving the sampled within-genome evidence insufficient to establish a shared ancestral WGD in Boraginaceae. Testing shared ancestry would require cross-species collinearity or phylogenomic duplication mapping that places homologous duplicated blocks or gene duplications on the ancestral branch. The Cordiaceae, Ehretiaceae and Hydrophyllaceae results each derive from one genome and therefore apply only to the sampled lineage. The *E. macrophylla* component also shows why absolute-age comparisons must be cautious: applying different assumed rates to the same K_s value shifts the age algebraically, while different species and paralog-construction methods add further non-equivalence.

For *N. menziesii*, assembly incompleteness alone is not an adequate explanation: genome-mode BUSCO completeness was 93.8% and the assembly is scaffold-level. However, protein-mode completeness was lower (82.3%), the genome was annotated de novo without a public reference annotation, and its number of self-synteny anchor pairs was the lowest of the seven. A genuine absence of a

younger lineage-specific duplication would produce a similarly sparse young-duplicate signal, while K_s saturation, alignment filtering and the 60.3% fitting-window movement further reduce resolution. One genome cannot separate these possibilities. Chromosome-scale assemblies for additional Hydrophylloideae genera, sampling of several species within the subfamily, and chromosome counts or cytogenetic evidence are needed; until then, no family-wide absence or presence claim is made.

Contrasting dynamics of TE load and chromosomal architecture: Gene-based macro-synteny remained detectable across assemblies that differ 11.5-fold in size. We could not test whether TE accumulation was concentrated in pericentromeric regions or dispersed along chromosome arms because several assemblies lack chromosome-scale pseudomolecules and comparable centromere annotations. The preserved collinearity describes gene order; its relationship to the chromosomal distribution of TE insertions remains unresolved.

Limitations: The seven assemblies are phylogenetically non-independent, and Cordiaceae, Ehretiaceae and Hydrophyllaceae are each represented by one genome. Assembly and annotation provenance differ among species. K_s mixtures and algebraic age conversions depend on input construction, model selection and assumed substitution rates; no alternative clustering family or lineage-specific rate test was performed. Complete command records are unavailable for some cross-species MCscan comparisons. The merged-CDS/EDTA-overlap analysis is a positional sensitivity check, not a functional identification of TE-derived gene models or a direct quantification of spurious anchors. Uniform centromere annotations and cytogenetic validation are unavailable. These limitations restrict causal, family-wide and precise chronological interpretation.

Conclusion

Across the seven sampled genomes, TE content varied from 13.31% to 76.81% of non-N sequence and was strongly associated with assembly size. Differences in TE composition and identity-derived intact-LTR age distributions document heterogeneous TE landscapes across Boraginales, but they do not resolve complete TE turnover, causality or adaptive significance. Two sampled genomes contained anchor-supported, window-stable paleoduplication candidate components, while the other five screened components remained descriptive. These within-genome signals do not establish a shared ancestral event or yield precise WGD dates. Gene-based macro-synteny remained detectable across the four sampled families despite broad differences in analysed sequence-set size and TE abundance. Broader chromosome-scale sampling and cytogenetic evidence are needed, particularly within Hydrophylloideae.

Acknowledgements

We acknowledge the genome initiatives, research teams and previous work that generated and released the Boraginales genome resources used in this study, including the previously released *Echium plantagineum* assembly reported by Tang *et al*., (2020), the Purdue *Lithospermum erythrorhizon* genome project, the Darwin Tree of Life project at the Wellcome Sanger Institute, the California Conservation Genomics Project, and the authors of the published *Cordia subcordata*, *Ehretia macrophylla* and *Pectocarya recurvata* genome resources. This research was supported by the National Natural Science Foundation of China (U1903201, 31771413), the State Key Laboratory of Tree Genetics and Breeding (SKLTGB-NJ2024-101), and the Program for Changjiang Scholars and Innovative Research Team in University from the Ministry of Education of China (IRT_14R27).

Author's Contribution: Wencai Jie conceived the study with Guihua Lu and Yonghua Yang, designed and performed the comparative bioinformatic analyses, interpreted the results and drafted the manuscript. Yonghui Liao, Shoucheng Huang and Xiaohui Lai collected and curated genomic and transcriptomic datasets and contributed to data validation and result interpretation. Yuxin Zhang, Xiaoran Lv, Changyi Wang, Minkai Yang, Zhongling Wen, Hongwei Han and Jinliang Qi curated data and literature, performed quality control and validated the results. Tongming Yin provided resources and methodological advice. Guihua Lu and Yonghua Yang supervised and administered the study and provided resources; Guihua Lu contributed to interpretation, and Yonghua Yang acquired funding. All authors reviewed and revised the manuscript and approved the final version.

Conflict of Interest: The authors declare that there is no conflict of interest.

References

Auber, R.P., T. Suttiyut, R.M. McCoy, M. Ghaste, J.W. Crook, A.L. Pendleton, J.R. Widhalm and J.H. Wisecaver. 2020. Hybrid de novo genome assembly of red gromwell (*Lithospermum erythrorhizon*) reveals evolutionary insight into shikonin biosynthesis. *Hort. Res.*, 7(1): 82.

Buchfink, B., K. Reuter and H.-G. Drost. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*, 18(4): 366-368.

Chen, Y.-L., Z.-F. Wang, S.-G. Jian, H.-M. Liao and D.-M. Liu. 2023. Genome Assembly of *Cordia subcordata*, a Coastal Protection Species in Tropical Coral Islands. *Int. J. Mol. Sci.*, 24(22): 16273.

Cheng, S., Q. Zhang, X. Geng, L. Xie, M. Chen, S. Jiao, S. Qi, P. Yao, M. Lu, M. Zhang, W. Zhai, Q. Yun and S. Feng. 2024. Haplotype-resolved chromosome-level genome assembly of *Ehretia macrophylla*. *Sci. Data*, 11(1): 589.

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

Gaut, B.S., B.R. Morton, B.C. McCaig and M.T. Clegg. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc. Natl. Acad. Sci. U.S.A.*, 93(19): 10274-10279.

Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman and A. Regev. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7): 644-652.

Haas, B.J., S.L. Salzberg, W. Zhu, M. Pertea, J.E. Allen, J. Orvis, O. White, C.R. Buell and J.R. Wortman. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.*, 9(1): R7.

Hoang, D.T., O. Chernomor, A. von Haeseler, B.Q. Minh and L.S. Vinh. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.*, 35(2): 518-522.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A.F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez and S. Hunter. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9): 1236-1240.

Katoh, K. and D.M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4): 772-780.

Keilwagen, J., F. Hartung and J. Grau. 2019. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. In: M. Kollmar (ed.), *Gene Prediction: Methods and Protocols*. Methods in Molecular Biology, vol. 1962, pp. 161-177. Humana Press, New York.

Keilwagen, J., F. Hartung, M. Paulini, S.O. Twardziok and J. Grau. 2018. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinform.*, 19(1): 189.

Kiełbasa, S.M., R. Wan, K. Sato, P. Horton and M.C. Frith. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res.*, 21(3): 487-493.

Landis, J.B., D.E. Soltis, Z. Li, H.E. Marx, M.S. Barker, D.C. Tank and P.S. Soltis. 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. *Amer. J. Bot.*, 105(3): 348-363.

Lee, S.-I. and N.-S. Kim. 2014. Transposable Elements and Genome Size Variations in Plants. *Genom. Inform.*, 12(3): 87-97.

Luebert, F., L. Cecchi, M.W. Frohlich, M. Gottschling, C.M. Guilliams, K.E. Hasenstab-Lehman, H.H. Hilger, J.S. Miller, M. Mittelbach, M. Nazaire, M. Nepi, D. Nocentini, D. Ober, R.G. Olmstead, F. Selvi, M.G. Simpson, K. Sutorý, B. Valdés, G.K. Walden and M. Weigend. 2016. Familial classification of the Boraginales. *Taxon*, 65(3): 502-522.

Ma, J. and J.L. Bennetzen. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. U.S.A.*, 101(34): 12404-12410.

Manni, M., M.R. Berkeley, M. Seppey, F.A. Simão and E.M. Zdobnov. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.*, 38(10): 4647-4654.

Michael, T.P. 2014. Plant genome size variation: bloating and purging DNA. *Brief. Funct. Genom.*, 13(4): 308-317.

Northing, P.C., J.A. Pelosi, D.L. Venable and K.M. Dlugosch. 2025. Chromosome-scale reference genome of *Pectocarya recurvata*, the species with the smallest reported genome size in Boraginaceae. *Appl. Plant Sci.*, 13(3): e70008.

Ou, S., N. Jiang, C.N. Hirsch and M.B. Hufford. 2024. Response to Commentary: Accounting for diverse transposable element landscapes is key to developing and evaluating accurate de novo annotation strategies. *Genome Biol.*, 25(1): 6.

Ou, S., W. Su, Y. Liao, K. Chougule, J.R.A. Agda, A.J. Hellinga, C.S.B. Lugo, T.A. Elliott, D. Ware, T. Peterson, N. Jiang, C.N. Hirsch and M.B. Hufford. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.*, 20(1): 275.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay. 2011. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, 12(85): 2825-2830.

Pertea, M., G.M. Pertea, C.M. Antonescu, T.-C. Chang, J.T. Mendell and S.L. Salzberg. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, 33(3): 290-295.

Scott, D.W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York.

Shaffer, H.B., E. Toffelmier, R.B. Corbett-Detig, M. Escalona, B. Erickson, P. Fiedler, M. Gold, R.J. Harrigan, S. Hodges, T.K. Luckau, C. Miller, D.R. Oliveira, K.E. Shaffer, B. Shapiro, V.L. Sork and I.J. Wang. 2022. Landscape Genomics to Enable Conservation Actions: The California Conservation Genomics Project. *J. Hered.*, 113(6): 577-588.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack and B. Morgenstern. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucl. Acids Res.*, 34(Web Server): W435-W439.

Steenwyk, J.L., T.J. Buida III, Y. Li, X.-X. Shen and A. Rokas. 2020. ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.*, 18(12): e3001007.

Tang, C.-Y., S. Li, Y.-T. Wang and X. Wang. 2020. Comparative genome/transcriptome analysis probes Boraginales' phylogenetic position, WGDs in Boraginales, and key enzyme genes in the alkannin/shikonin core pathway. *Mol. Ecol. Resour.*, 20(1): 228-241.

Tang, H., J.E. Bowers, X. Wang, R. Ming, M. Alam and A.H. Paterson. 2008. Synteny and Collinearity in Plant Genomes. *Science*, 320(5875): 486-488.

Tang, H., V. Krishnakumar, X. Zeng, Z. Xu, A. Taranto, J.S. Lomas, Y. Zhang, Y. Huang, Y. Wang, W.C. Yim, J. Zhang and X. Zhang. 2024. JCVI: A versatile toolkit for comparative genomics analysis. *iMeta*, 3(4): e211.

Tegenfeldt, F., D. Kuznetsov, M. Manni, M. Berkeley, E.M. Zdobnov and E.V. Kriventseva. 2025. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucl. Acids Res.*, 53(D1): D516-D522.

Vasile, M.-A., T. Böhnert, J. Jeiter, D. Cardoso, P.W. Moonlight and M. Weigend. 2025. An updated phylogeny of Boraginales based on the Angiosperms353 probe set: a roadmap for understanding morphological evolution. *Ann. Bot.*, 136(1): 77-97.

Virtanen, P., R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, 17(3): 261-272.

Wang, D., Z. Zheng, Y. Li, H. Hu, Z. Wang, X. Du, S. Zhang, M. Zhu, L. Dong, G. Ren and Y. Yang. 2021. Which factors contribute most to genome size variation within angiosperms? *Ecol. Evol.*, 11(6): 2660-2668.

Wong, T.K.F., N. Ly-Trong, H. Ren, P. Demotte, H. Baños, A.J. Roger, E. Susko, C. Bielow, N. De Maio, N. Goldman, M.W. Hahn, M. dos Reis, L.S. Vinh, G. Huttley, R. Lanfear and B.Q. Minh. 2026. IQ-TREE 3: Phylogenomic Inference Software Using Complex Evolutionary Models. *Mol. Biol. Evol.*, 43(5): msag117.

Zhang, C., T. Zhang, F. Luebert, Y. Xiang, C.-H. Huang, Y. Hu, M. Rees, M.W. Frohlich, J. Qi, M. Weigend and H. Ma. 2020. Asterid Phylogenomics/ Phylotranscriptomics Uncover Morphological Evolutionary Histories and Support Phylogenetic Placement for Numerous Whole-Genome Duplications. *Mol. Biol. Evol.*, 37(11): 3188-3210.

Zwaenepoel, A. and Y. Van de Peer. 2019. wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics*, 35(12): 2153-2155.