scientific reports

Check for updates

OPEN

Urticaceae542 as a novel probe set combining universal and lineage-specific loci for phylogenomics of Urticaceae

Han-Chun Chien & Yu-Hsin Tseng✉

Hybridization-based target enrichment sequencing (Hyb-Seq) is a powerful approach for phylogenomic studies of non-model organisms, particularly those with complex genomes or from degraded materials such as herbarium specimens. To improve phylogenetic resolution within the challenging Urticaceae family, we developed a new probe set, Urticaceae542, by combining the universal Angiosperms353 loci with lineage-specific loci from three representative Urticaceae genera. After evaluating its performance across 61 Urticaceae species, including herbarium specimens, we found that Urticaceae542 effectively facilitated both deep and shallow-scale phylogenomic analyses. The lineage-specific loci (Urti-MarkerMiner) provided more parsimony-informative sites and longer sequences, providing greater resolution for recent divergence events. In contrast, the Angiosperms353-derived loci (Urti-Angiosperms353) offered higher locus recovery efficiency and reduced missing data, features crucial for robust deep-level phylogenetic relationships. Moreover, the performance of Urticaceae542 was not significantly affected by sample age or preservation condition, but its efficiency varied significantly across different taxonomic clades and genera. Our findings demonstrate that combining universal and taxon-specific probes generates a robust strategy for resolving complex evolutionary histories in challenging plant families, and the Urticaceae542 provides a powerful new resource for phylogenomics of Urticaceae.

Keywords Angiosperms353, Hyb-Seq, Phylogenomics, Taxon-specific probes, Urticaceae542

Hybridization-based target enrichment sequencing (Hyb-Seq) is a powerful genomic approach that combines targeted capture of low-copy nuclear genes with genome skimming to simultaneously hundreds to thousands of nuclear loci[1]. It has proven especially useful for phylogenomic studies of non-model organisms, taxa with complex genomes, and degraded samples such as herbarium specimens[2,3]. The method offers high efficiency and cost-effectiveness, yielding large-scale sequence datasets across broad-scale phylogenetics (e.g., Annonaceae[4] and Compositae[5], to fine-scale population genomics (e.g., *Philadelphus hitchcockianus* and *Salvia summa*)[6] to study deep phylogenetic relationships to analyzing recent speciation events and genetic diversity in conservation genomics.

A major advancement in Hyb-Seq methodology has been the development of universal probe sets, which enable consistent and comparable data collection across diverse taxa. One of the most widely adopted universal probe kits is Angiosperms353, which targets 353 putative single-copy nuclear genes conserved across flowering plants[7]. However, this universal probe kit reliance on highly conserved regions may constrain their ability to resolve the relationships at shallower taxonomic scales, particularly in rapidly radiating lineages, which are often characterized by incomplete lineage sorting, historical introgression, or low genetic divergence[8].

Studies comparing Angiosperms353 to lineage-specific probes have shown that taxon-specific probe sets generally recover more loci, longer alignments, and higher parsimony-informative site counts, especially for shallow-scale phylogenies[9]. The recent development bioinformatic tools—such as MarkerMiner[10] and LoCoLotive[11]—has enabled the development of customized probe sets to facilitate the identification of low-copy, orthologous nuclear loci from transcriptome or reference genome, which largely decrease the bioinformatic effort. However, taxon-specific probe sets tend not to be readily combinable with data generated using other probe sets, leading to higher cost per sample[12]. A combination of universal and taxon-specific probes in a single

Department of Life Sciences, National Chung Hsing University, Taichung 40227, Taiwan. ✉email: yuhsin.tseng@nchu.edu.tw

capture reaction has been demonstrated to enhance locus recovery across taxonomic scales, enabling both shallow and deep-scale phylogenomic analyses and facilitates data integration with existing datasets[13,14].

Target capture efficiency in Hyb-Seq can be influenced by various factors, including probe design (sequence similarity, probe length, representative diversity)[7,12], genome size[11,15], taxonomic level[16,17], and DNA quality and quantity (DNA input, library quality/complexity)[6,18]. The sample preservation method (sample age, herbarium/silica-dried)[19,20] is particularly relevant to DNA quality. Herbarium and silica-dried specimens are an irreplaceable source of verified plant material, especially for rare or hard-to-collect taxa. Despite potential DNA degradation in these historical samples, target capture approaches effectively recover phylogenetically informative loci, allowing their integration into modern genomic studies with minimal compromise to data quality[19,20]. Therefore, when designing new probe sets, the ability to successfully recover data from herbarium materials becomes especially critical for expanding taxonomic and temporal sampling.

The family Urticaceae, comprising over 2,000 species across 59 genera[21], presents unique phylogenetic challenges that underscore its suitability for a Hyb-Seq approach. Though understanding evolutionary relationships within this diverse lineage is crucial, comprehensive phylogenetic analyses of Urticaceae are remarkably scarce, with only three studies published to date[22–24]. The most current molecular phylogeny, established by Wu et al.[23], delineated Urticaceae into four well-supported clades (Clade I, II, III & IV). While homoplasy in the morphological characters limits the capacity to circumscribe these major clades solely through synapomorphies, combined suites of characters, such as the shape and presence of cystoliths, connate pistillate perianth, and presence of stinging hairs, enable their morphological diagnosis[25]. However, these existing efforts have not fully resolved the evolutionary history of the family. A critical limitation lies in the limited taxon sampling and/or insufficient genomic data coverage in these prior investigations, leaving numerous phylogenetic relationships ambiguous. This is particularly apparent in the unresolved tribal delimitations, the uncertain phylogenetic placement of several genera (e.g., *Cecropia*, *Chamabainia*, *Coussapoa*), and the unclear circumscription of certain polyphyletic genera (e.g., *Boehmeria*, *Laportea*, *Pouzolzia*, and *Urera*)[23].

Further compounding these difficulties, Urticaceae tissues frequently contain high concentrations in polysaccharides and polyphenols[26], compounds known to inhibit DNA extraction and subsequent enzymatic reactions. Given that Hyb-Seq can effectively utilize degraded DNA from challenging sources, this methodological flexibility makes it particularly well-suited for addressing the complex phylogenetic questions within Urticaceae. The first phylogenomic analysis of Urticaceae using the Angiosperms353 probe set was conducted by Baker et al.[27], including 73 species representing 52 genera. While the study appears to provide a robust phylogenetic framework at the generic level, the performance of Angiosperms353 probe set— specifically its capture efficiency, paralog detection, and species delimitation capabilities—has yet to be systematically evaluated within Urticaceae.

To address these recognized gaps in Urticaceae phylogenetics, we developed Urticaceae542, a novel probe set specifically targeting 542 nuclear loci within the family. This study therefore aims to achieve these key objectives: (1) Evaluate the performance of Urticaceae542 across various clades and genera, comparing its efficiency with both universal (Angiosperms353) and Urticaceae-specific probe sets. (2) Assess the impact of sample attributes—with a particular focus on sample age and herbarium specimen—on locus recovery and assembly quality. (3) Demonstrate the utility of Urticaceae542 in accurately reconstructing phylogenies and significantly enhancing genomic resolution across the Urticaceae.

Result

Urticaceae542 probe design

Following the MarkerMiner analysis and subsequent filtering steps, a total of 298 loci were identified as candidate low-copy loci (designated as "Urti-MarkerMiner" probe set in this study). Another 244 Angiosperms353 orthologs derived from *Boehmeria nivea* and *Elatostema rivulare* were added to the candidate loci list ("Urti-Angiosperms353" probe set). Twenty-eight loci were shared between two initial probe sets. After further filtering by the Arbor Biosciences team, the final probe set was constructed from 542 loci, merged and named as a "Urticaceae542" probe set. The final and filtered probe set consisted of 39,862 probes.

Sequencing information and recovered loci

Results of sequencing and assembly statistics are summarized in Table 1; Figs. 1 and 2 and Supplementary Table S1. The total number of trimmed reads for all Urticaceae samples ranged from 2,404,954 (*Leucosyke quadrinervia*) to 23,905,374 (*Gonostegia hirta*) with mean of 7,984,493.541. The percentage of on-target reads varied from 18.4% in *Pilea aquarum* ssp. *brevicornuta* to 82.9% in *Boehmeria nivea* var. *tenacissima* (mean = 56.7%), which related to the sequence depth varied from the mean of 22 ± 52 in *Pilea aquarum* ssp. *brevicornuta* to 1,209 ± 1,912 in *Boehmeria nivea* var. *tenacissima* (mean = 281.5) (Pearson $r = 0.661$, $p < 0.01$). The number of recovered loci varied from 476 in *Laportea aestuans* to 540 in *Procris montana*, with an average of 521.3 loci recovered.

Assembly comparison among clades and genera

In the comparison of sequencing and assembly results among different clade, percentage of on-target read, number of target loci recovered, sequence length (including exon and supercontig) showed statistically significant differences across major clades (all $p < 0.01$) (Table 1; Fig. 1). The definition of four major clades (Clade I, II, III, IV) was based on classification system proposed by Wu et al.[23] (Fig. 4). Clade I consistently showed the highest mean values across the highest mean percentage of on-target reads (66.0%) (Fig. 1a), the mean of exon (657,755 bp) (Fig. 1c) and supercontig length (1,561,242 bp) (Fig. 1d). Clade IV showed the lowest on-target reads (45.2%) (Fig. 1a) but with the highest number of target loci recovered (532.5) (Fig. 1b). However, these values for Clade IV should be interpreted with caution due to the limited sample size ($n = 2$), which may not adequately represent the clade's true variation. In addition, Clade III showed the lowest averages

Clade[a]	Genus	Species	Total reads[b]	On target reads (%)[c]	Total targets[d]	Exon length (bp)[e]	Supercontig length (bp)[f]	Mean target depth ± SD	Paralogs[g]	Sample type[h]	Coll. Year
Clade I	*Boehmeria*	*Boehmeria clidemioides*	12,475,634	66.7	530	667,764	1,548,475	462 ± 969	6	Fresh	2023
		Boehmeria. formosana	6,877,218	65.4	529	666,342	1,558,584	276 ± 515	4	Fresh	2023
		Boehmeria hwaliensis	12,663,656	69.7	528	666,675	1,549,614	481 ± 1076	8	Fresh	2023
		Boehmeria longispica	14,165,904	69.5	530	660,558	1,435,132	614 ± 1195	2	Fresh	2023
		Boehmeria nivea	19,951,776	80.2	533	728,679	1,722,997	1138 ± 1730	8	Fresh	2022
		Boehmeria nivea var. *tenacissima*	21,367,918	82.9	534	728,202	1,702,894	1209 ± 1912	10	Fresh	2022
		Boehmeria pilosiuscula	6,201,356	67.1	531	666,237	1,547,229	256 ± 502	3	Fresh	2022
		Boehmeria pilushanensis	12,578,050	67.6	529	665,922	1,549,658	463 ± 993	4	Fresh	2022
		Boehmeria zollingeriana var. *podocarpa*	7,871,370	67	534	687,852	1,595,669	328 ± 603	142	Fresh	2023
	Chamabainia	*Chamabainia cuspidata*	7,996,100	69.3	524	657,345	1,552,150	347 ± 644	2	Fresh	2022
	Cypholophus	*Cypholophus moluccanus*	7,448,934	47.3	533	653,766	1,481,419	221 ± 422	2	Fresh	2023
	Debregeasia	*Debregeasia orientalis*	15,748,254	78.1	533	720,963	1,688,510	807 ± 1338	3	Fresh	2022
	Gonostegia	*Gonostegia hirta*	23,905,374	53.9	522	599,661	1,363,706	745 ± 1641	6	Fresh	2022
		Gonostegia matsudae	8,126,970	62.1	509	586,134	1,259,487	294 ± 677	3	Fresh	2023
		Gonostegia pentandra	9,671,792	62.7	506	577,299	1,262,874	126 ± 786	0	Fresh	2023
	Oreocnide	*Oreocnide pedunculata*	5,449,532	64	534	681,291	1,683,202	213 ± 417	4	Fresh	2022
		Oreocnide trinervis	4,490,350	41.9	538	666,336	1,465,904	109 ± 260	4	Fresh	2023
	Pipturus	*Pipturus arborescens*	6,474,626	71	533	670,929	1,518,580	301 ± 588	2	Fresh	2023
	Pouzolzia	*Pouzolzia taiwaniana*	12,914,836	67.6	515	612,822	1,501,332	457 ± 1482	2	Fresh	2023
		Pouzolzia zeylanica	6,000,418	65.7	510	590,328	1,400,122	232 ± 538	5	Fresh	2022
Clade II	*Elatostema*	*Elatostema edule*	18,715,880	59.6	537	644,523	1,397,679	745 ± 1326	10	Fresh	2023
		Elatostema goudotianum	7,469,036	64.3	535	643,530	1,524,405	282 ± 627	11	Herbarium (BM)	1992
		Elatostema grijsii	2,935,500	54.5	533	646,599	1,503,363	94 ± 187	15	Fresh	2010
		Elatostema hypoglaucum	4,713,492	70.3	533	667,584	1,442,932	220 ± 382	11	Fresh	2022
		Elatostema incisum	9,389,212	68.3	537	634,761	1,399,196	384 ± 1196	12	Herbarium (MO)	1993
		Elatostema madagascariense	8,518,504	61.5	537	635,490	1,381,470	312 ± 879	11	Herbarium (MO)	2007
		Elatostema nasutum	12,307,094	62.7	531	637,608	1,410,806	489 ± 979	14	Fresh	na
		Elatostema parvum	2,548,656	61	533	631,737	1,413,202	95 ± 187	4	Fresh	2022
		Elatostema rugosum	11,516,870	66.6	532	662,625	1,355,836	500 ± 893	6	Herbarium (BM)	na
		Elatostema sinense var. longicornutum	9,900,858	61.3	538	640,125	1,429,054	365 ± 718	7	Herbarium (BM)	2005
		Elatostema subcoriaceum	5,529,586	71.6	534	667,449	1,438,178	94 ± 526	9	Fresh	2023
		Elatostema suzukii	11,256,430	68.5	530	665,859	1,516,686	509 ± 875	18	Herbarium (K)	1947
		Elatostema viridis	3,737,502	59.7	538	641,565	1,325,027	131 ± 265	11	Herbarium (MO)	1999
		Elatostema welwitschii	5,177,068	80	535	673,518	1,418,811	273 ± 472	11	Herbarium (BM)	1855
	Elatostematoides	*Elatostematoides filicoides*	5,439,160	33.9	538	603,930	1,281,859	98 ± 392	2	Herbarium (BM)	1946
		Elatostematoides fruticulosa	5,005,494	44.3	535	618,129	1,398,717	121 ± 278	4	Herbarium (BM)	1932
	Pilea	*Pilea angulata*	4,897,514	38.5	500	484,440	1,204,649	92 ± 238	3	Fresh	2022
		Pilea aquarum subsp. *brevicornuta*	2,548,820	18.4	518	448,227	1,020,885	22 ± 52	5	Fresh	2022
		Pilea elliptifolia	4,936,556	38.6	503	497,301	1,257,482	94 ± 224	5	Fresh	2022
		Pilea funkikensis	5,278,512	38.3	503	488,370	1,209,561	98 ± 266	4	Fresh	2022
		Pilea matsudae	3,491,220	39.7	504	478,485	1,251,910	68 ± 178	5	Fresh	2022
		Pilea melastomoides	4,448,690	44.2	494	427,125	1,136,642	93 ± 237	2	Fresh	2022
		Pilea microphylla	6,399,624	42.6	494	413,430	1,135,351	94 ± 339	38	Fresh	2023
		Pilea plataniflora	13,113,356	63.2	498	471,390	1,164,366	377 ± 1271	3	Fresh	2022
		Pilea pumila	6,628,468	48.6	511	465,795	1,187,562	153 ± 370	14	Fresh	2022
		Pilea rotundinucula	5,193,322	29.4	511	471,216	1,153,100	70 ± 186	5	Fresh	2022
		Pilea somae	3,494,366	36.7	501	471,783	1,203,406	62 ± 153	3	Fresh	2022
		Pilea peploides	5,419,368	44	508	460,797	1,150,757	110 ± 289	73	Fresh	2023
	Procris	*Procris frutescens*	8,490,996	52.3	533	619,464	1,481,796	242 ± 561	22	Herbarium (K)	1984
		Procris laevigata	5,455,682	53.5	537	611,949	1,456,298	162 ± 382	7	Fresh	2022
		Procris montana	4,408,510	60.7	540	605,631	1,318,981	150 ± 318	7	Herbarium (MO)	1989
Clade III	*Dendrocnide*	*Dendrocnide kotoensis*	4,317,376	57.7	520	571,065	1,269,246	129 ± 294	46	Fresh	2023

Continued

Clade[a]	Genus	Species	Total reads[b]	On target reads (%)[c]	Total targets[d]	Exon length (bp)[e]	Supercontig length (bp)[f]	Mean target depth ± SD	Paralogs[g]	Sample type[h]	Coll. Year
		Dendrocnide meyeniana	3,115,384	54.4	516	573,471	1,392,257	88 ± 215	69	Fresh	2022
	Girardinia	*Girardinia diversifolia*	4,548,978	53.4	519	549,963	1,260,110	126 ± 265	1	Fresh	2022
	Laportea	*Laportea aestuans*	5,555,026	47.2	476	451,602	1,145,997	121 ± 337	3	Fresh	2023
		Laportea interrupta	8,465,672	43.4	482	457,551	1,171,370	178 ± 448	2	Fresh	2022
	Nanocnide	*Nanocnide japonica*	5,987,368	53.6	491	475,257	1,113,377	147 ± 462	3	Fresh	2022
	Urtica	*Urtica taiwaniana*	6,960,822	45.6	502	597,282	1,274,688	194 ± 473	77	Fresh	2022
		Urtica thunbergiana	8,322,546	54.4	485	585,393	1,367,336	256 ± 595	5	Fresh	2022
Clade IV	*Leucosyke*	*Leucosyke quadrinervia*	2,404,954	38.4	532	597,327	1,301,971	51 ± 133	0	Fresh	2022
	Maoutia	*Maoutia setosa*	6,630,566	51.9	533	623,799	1,417,228	201 ± 427	3	Fresh	2023
outgroup	*Ficus*	*Ficus carica*	54,332,174	1.1	483	424,185	1,039,956	15 ± 924	0	NCBI (DRR498457)	

Table 1. Summary of sequencing and assembly statistics for urticaceae samples enriched using the urticaceae542 probe set. [a] Clade designation followed by the classification Wu et al.[23] [b] Total number of sequencing reads retained after adapter trimming and quality filtering. [c] Percentage of reads mapped to targets loci. [d] Number of target loci for which coding sequence were successfully assembled. [e] Total length of exon sequence assembled by HybPiper. [f] Total length of supercontigs (exon + intron) assembled by HybPiper. [g] Total number of putative paralogous loci recovered per sample. [h] Sample source: collected from fresh materials, herbarium specimen (herbarium code) or NCBI (accession number).

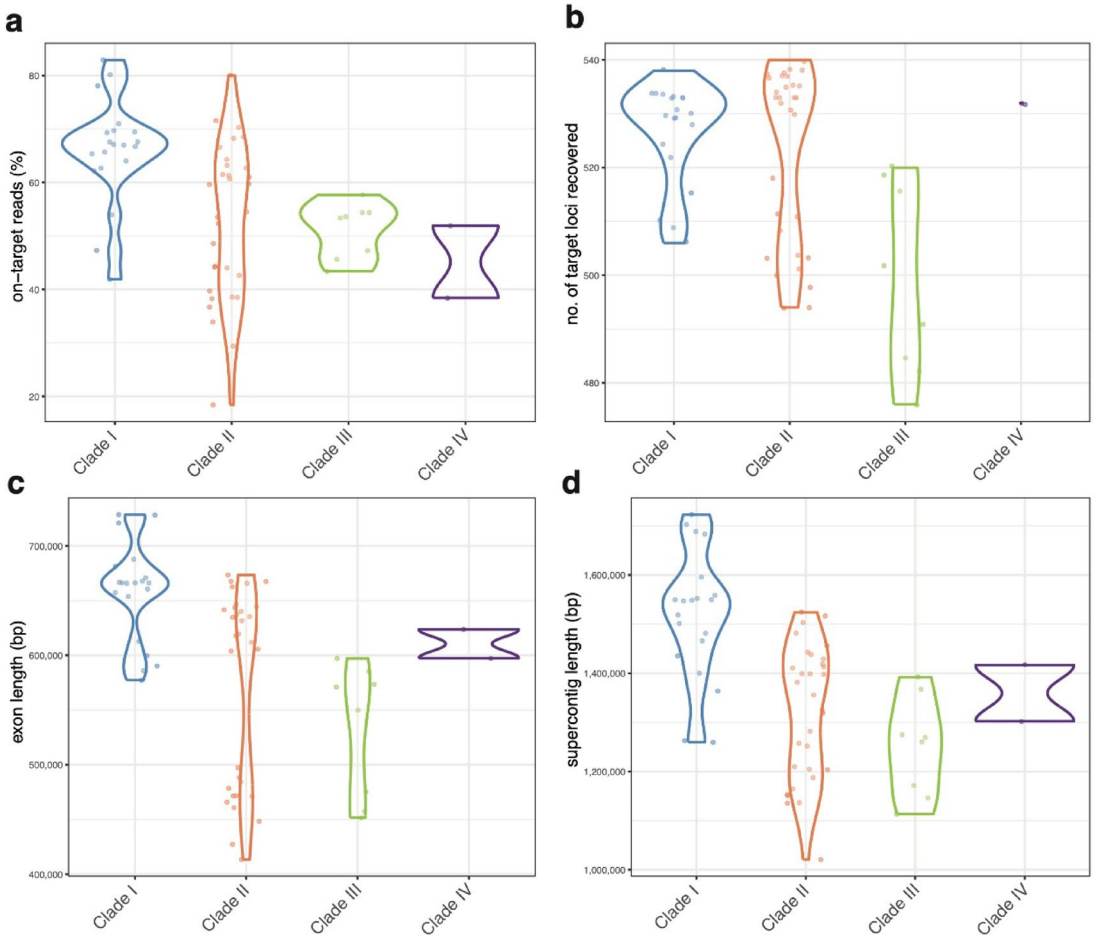

Fig. 1. Comparison of Urticaceae542 probe set performance in four major clades of Urticaceae. The figure illustrates the capture efficiency of Urticaceae542 probe set. (**a**) the percentage of on-target reads, (**b**) the number of target loci recovered, (**c**) the exon sequence length and (**d**) the supercontig sequence length. The four clades are designated based on Wu et al.[23].

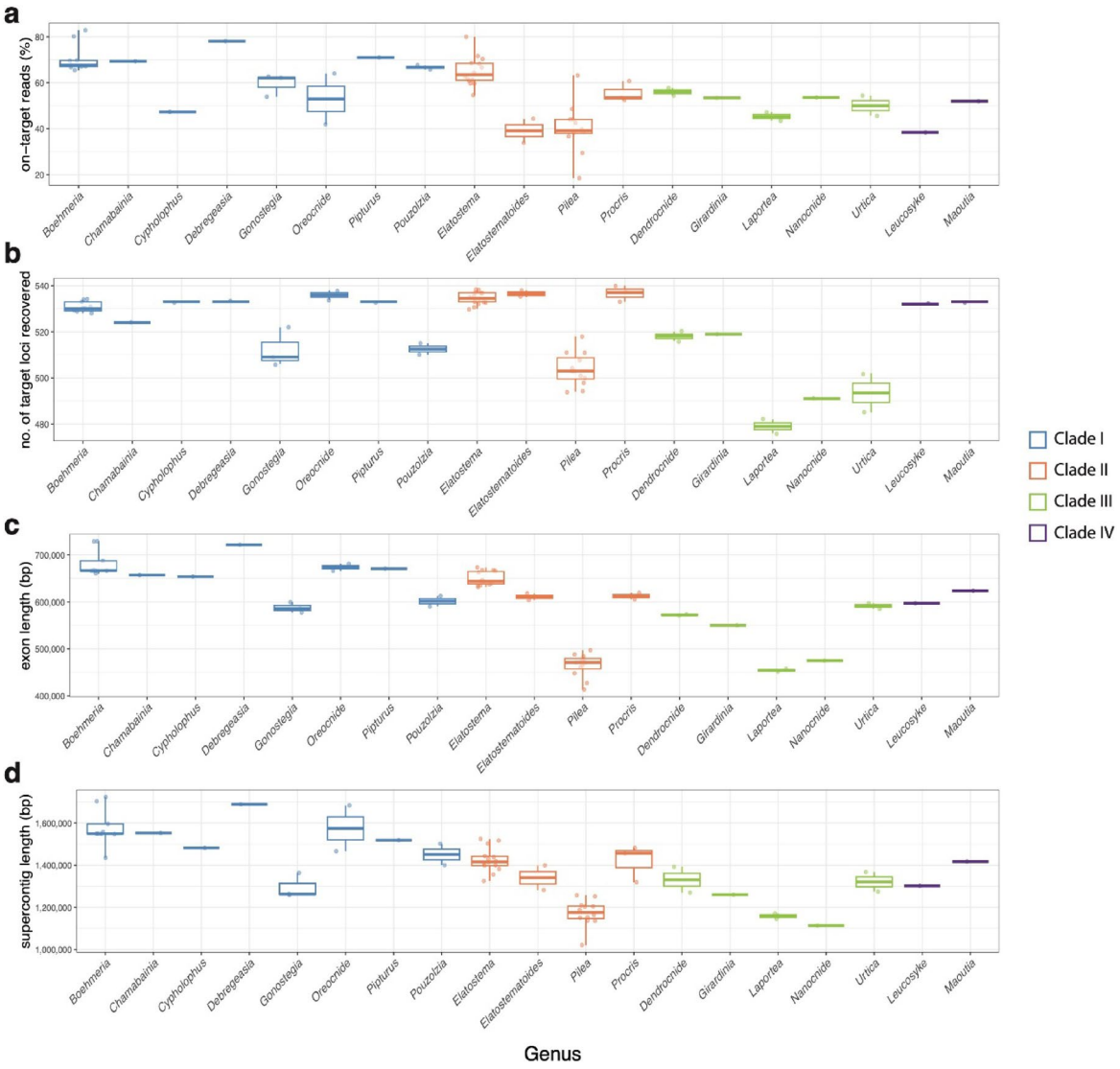

Fig. 2. Comparison of Urticaceae542 probe set performance across genera of Urticaceae. (**a**) the percentage of on-target reads, (**b**) the number of target loci recovered, (**c**) the exon sequence length and (**d**) the supercontig sequence length. The four clades are designated based on Wu et al.[23].

in number of target loci recovered (498.9) (Fig. 1b), the mean of exon (532,698 bp) (Fig. 1c) and supercontig length (1,249,297.6 bp) (Fig. 1d).

The assessment of genus-level variation within each clade revealed several distinct patterns. In Clade I, genera exhibited significant differences across the number of recovered loci ($F = 12.31$, $p < 0.01$) (Fig. 2b) and supercontig length ($F = 4.05$, $p = 0.017$) (Fig. 2d). The genera in Clade II showed more widespread variations, differing significantly with percentage of on-target reads ($F = 15.34$, $p < 0.01$) (Fig. 2a), target loci recovered ($F = 30.04$, $p < 0.01$) (Fig. 2b), exon length ($F = 18.99$, $p < 0.01$) (Fig. 2c) and supercontig length ($F = 13.41$, $p < 0.01$) (Fig. 2d). Within Clade II, *Elatostema* consistently exhibited the highest average values for most metrics, including the highest percentage of on-target reads (mean = 65.0%, range: 54.5–80.0%) (Fig. 2a), longest exon sequence length (mean = 649,498 bp) (Fig. 2c), and greatest supercontig length (mean = 1,425,475 bp) (Fig. 2d). In contrast, *Pilea* consistently showed the lowest performance across most metrics, including average number of recovered loci in Clade II (mean = 503.4, range: 494–518) (Fig. 2b) and the mean of exon (Fig. 2c) and supercontig sequence length (Fig. 2) (Fig. 2d). *Elatostematoides*, despite a relatively low and narrow range of on-target read percentages (33.9–44.3%) (Fig. 2a), performed comparably to higher-performing genera in terms of target loci recovered (mean = 536.5) (Fig. 2b) and exon sequence length (mean = 611,030 bp) (Fig. 2c), similar to *Elatostema* and *Procris*. In addition, Clade III showed no significant genus-level differences for any metric ($p > 0.05$). Clade IV was excluded from the analysis due to an insufficient number of genera and replicates.

Overall, significant differences among genera were detected across examined sequencing and assembly metrics, including exon sequence length ($H = 33.1$, $p < 0.01$), supercontig length ($F = 52.97$, $p < 0.01$), on-target reads ($H = 27.70$, $p < 0.01$), and recovered loci ($F = 70.21$, $p < 0.01$), while no. of reads was marginally non-

significant ($H = 13.54$, $p = 0.056$) (Fig. 2a). Among all genera, *Debregeasia* exhibited the highest values of average percentage of on-target reads (78.1%) (Fig. 2a), exon sequence length (720,963 bp) (Fig. 2c), and supercontig length (1,688,510 bp) (Fig. 2d). However, these statistics were based on a single sample. *Procris* had the highest average number of target loci recovered (536.7) (Fig. 2b). In contrast, *Leucosyke* showed the lowest on-target read percentage (38.4%) (Fig. 2a), and *Laportea* had the lowest average exon sequence length (454,577 bp) (Fig. 2c) and number of recovered target loci (479.0) (Fig. 2b), respectively. *Nanocnide* had the shortest average supercontig length (1,113,377 bp) (Fig. 2d).

Sample type and age comparison within *Elatostema*.

Capture performance of Urticaceae542 within *Elatostema* was not significantly influenced by sample type (fresh or herbarium). No statistically significant differences were detected between two samples types across all measure meterics, including number of reads ($U = 10$, $p = 0.268$), exon length ($U = 22$, $p = 0.53$), supercontig length ($t = 0.346$, $p = 0.737$), on-target read percentage ($t = 0.673$, $p = 0.519$), or number of recovered loci ($U = 8.0$, $p = 0.137$) (Table 1).

Similarly, the capture performance was not significantly correlated with specimen collection year, which ranged from 1855 to 2023. Fresh samples were collected between 2010 and 2023, and herbarium samples between 1855 and 2007. Spearman's rank correlation coefficients indicated weak and non-significant associations for number of reads (ρ = -0.151, $p = 0.6397$), exon length ($\rho = -0.053$, $p = 0.871$), supercontig length ($\rho = -0.186$, $p = 0.563$), on-target read percentage ($\rho = -0.326$, $p = 0.301$), and number of recovered loci ($\rho = -0.082$, $p = 0.800$) (Table 1).

Comparison of Urti-Angiosperms353, Urti-MarkerMiner and Urticaceae542 probe sets.

The probe set of Urti-MarkerMiner yielded significantly higher average metric than Urti-Angiosperms353 for on-target read percentage (mean = 41.6% vs. 15.5%, $t = 8.91$, $p < 0.01$) (Fig. 3a), number of recovered loci (mean = 281.4 vs. 240.0, $t = 9.72$, $p < 0.01$) (Fig. 3b), exon sequence length (mean = 443,776 bp vs. 152,457 bp, $t = 11.34$, $p < 0.01$) (Fig. 3c), and supercontig length (mean = 883,023 bp vs. 499,423 bp, $t = 8.64$, $p < 0.01$) (Fig. 3d). These differences likely reflected the design characteristics of the probe sets: Urti-MarkerMiner included a greater number of target loci (281 vs. 240) (Fig. 2b) with longer average length (ca. 1,623 bp) (Fig. 3c and d), compared to the shorter loci in Urti-Angiosperms353 (640.4 bp) (Fig. 3c and d) (Table 2).

Significant differences were also observed in intergeneric variation between the two probe sets. Results revealed that Urti-MarkerMiner exhibited significantly greater variance among genera in all key metrics (Fig. 3). Specifically, the number of recovered loci showed higher variance under Urti-MarkerMiner ($F = 20.40$, $p < 0.01$) (Fig. 3b), as did exon sequence length ($F = 54.99$, $p < 0.01$) (Fig. 3c), on-target read percentage ($F = 31.39$, $p < 0.01$) (Fig. 3a), and supercontig length ($F = 7.24$, $p < 0.01$) (Fig. 2d). These results may indicate that while Urti-MarkerMiner was more sensitive to taxon-specific genomic characteristics; in contrast, Angiosperms353 showed more consistent performance across taxa.

A comprehensive comparison of the two probe sets, Urti-Angiosperms353 and Urti-MarkerMiner, revealed clear differences in capture efficiency, sequence recovery, and alignment information (Table 2). Although Urti-MarkerMiner targeted a greater number of loci (mean = 281.3 vs. 240.0) with longer average sequence lengths (ca. 1,623 bp vs. 640 bp), Urti-Angiosperms353 exhibited significantly higher recovery efficiency, which yielded a greater proportion of genes recovered at > 50% of target length (mean = 96% vs. 86%; t = 7.38, $p < 0.01$) and > 75% of target length (mean = 84 vs. 65; $t = 6.39$, $p < 0.01$). Significant differences were also observed in alignment completeness and phylogenetic informativeness. Urti-Angiosperms353 alignments contained fewer missing data (mean = 8.3 vs. 11.7, $t = -4.42$, $p < 0.01$); in contrast, Urti-MarkerMiner alignments contained a greater number of parsimony-informative sites (mean = 0.41 vs. 0.46, $t = -17.56$, $p < 0.01$).

Paralog recovery

Based on HybPiper's "length" detection criterion, 188 target loci from the Urti-Angiosperms353 probe set and 106 from the Urti-MarkerMiner probe set yielded at least one paralog assembly. No single locus was identified as paralogous across all sampled species. *Boehmeria zollingeriana* var. *podocarpa* consistently produced the highest number of paralog-flagged loci in both sets (94 loci in Urti-Angiosperms353 and 48 in Urti-MarkerMiner) (Supplementary Table S1). Overall, Urti-Angiosperms353 set generated significantly more loci with paralog warnings per sample than Urti-MarkerMiner set ($t = 3.12$, $p < 0.01$). The percentage of recovered loci identified as paralogous per sample varied from 0% to 38.5% (mean = 4.1%) for Urti-Angiosperms353 and from 0% to 16.1% (mean = 1.0%) for Urti-MarkerMiner. Compared the paralogs assemblies among genera, on average, *Dendrocnide* samples yielded the most paralog assemblies in Urti-Angiosperms353 (mean = 42.5), followed by *Urtica* (19.5) and *Pilea* (19.5), whereas in Urti-MarkerMiner, *Urtica* samples showed the highest average (22.0), followed by *Dendrocnide* (15.0) and *Boehmeria* (6.0) (Supplementary Table S1).

Phylogenomic analysis

Phylogenetics analyses reconstructed based on exon and supercontig sequences employing both ASTRAL and RAxML methods showed largely similar topologies (Fig. 4, Supplementary Figs. S1–S7) based on all probe sets (Urticaceae542 and its two sub-probe sets, Urti-Angiosperms353, and Urti-MarkerMiner). Our results supported the four major clades of Urticaceae identified by Wu et al.[23], with Clades I and IV forming a sister group (LPP = 1, BS = 100), and Clades II and III forming another (LPP = 1, BS = 100). Most sampled genera were recovered as monophyletic, including *Dendrocnide*, *Elatostema*, *Elatostematoides*, *Gonostegia*, *Laportea*, *Oreocnide*, *Pilea*, *Pouzoizia*, *Procris* and *Urtica*, while *Boehmeria* was a polyphyletic group.

Despite the overall congruence and high support value for most relationships, two key phylogenetic conflicts were identified based on the different probe sets and analytical methods. The first one involved the relationships among three major subclades within Clade I. In Clade I, *Oreocnide* species consistently appeared as sister to the remaining species, which formed three distinct clades: Clade I-1 (a monophyletic group comprising most

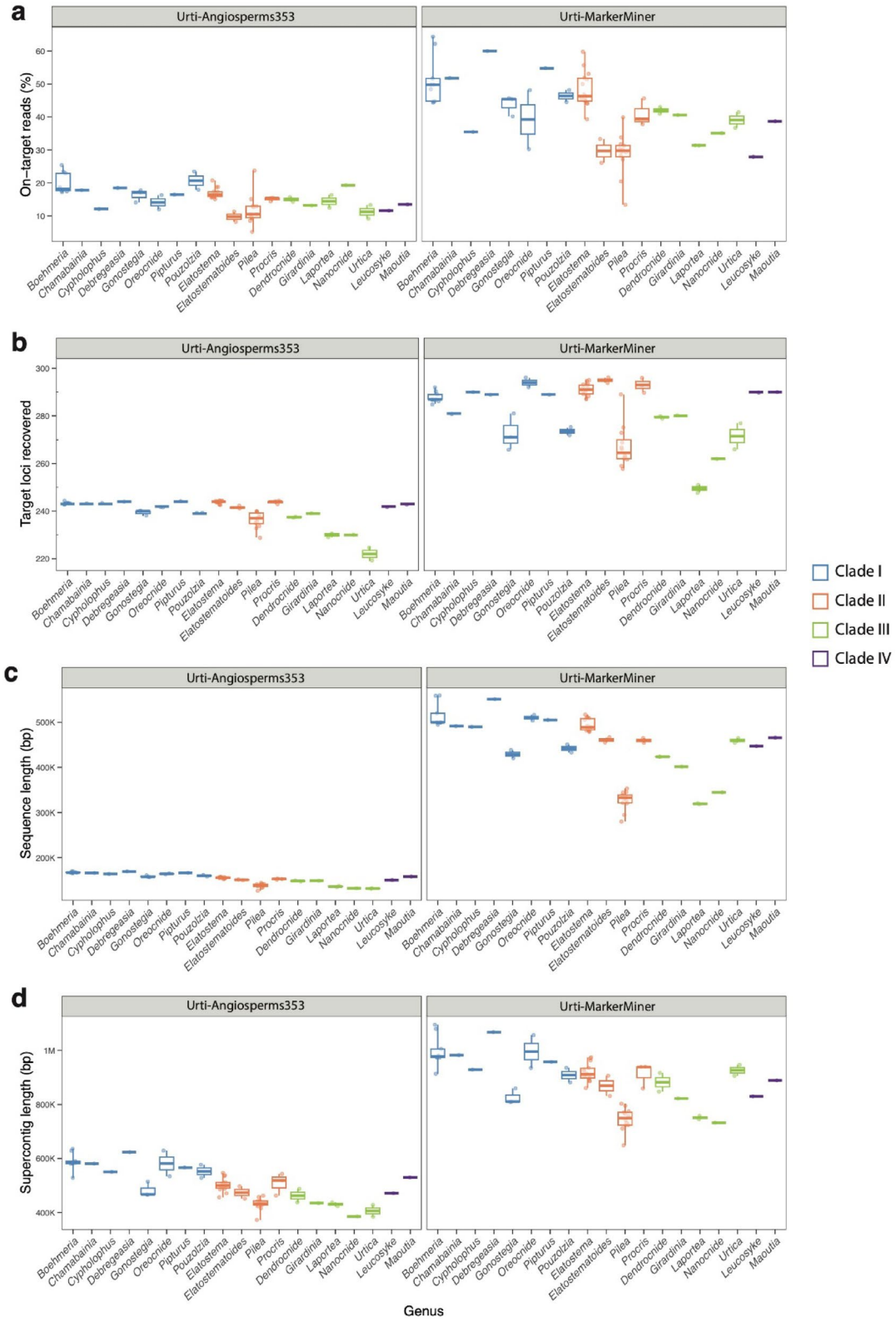

Fig. 3. Comparison of Urti-Angiosperms353 and Urti-MarkerMiner probe sets performance across genera of Urticaceae. (**a**) the percentage of on-target reads, (**b**) the number of target loci recovered, (**c**) the exon sequence length and (**d**) the supercontig sequence length. The four clades are designated based on Wu et al.[23].

	Urti-Angiosperms353	Urti-MarkerMiner	Urticaceae542
Recovered loci[a] / Total loci	240 / 244	281.3 / 298	521.3/542
Average recovered sequence length per locus (bp)	640.4	1,623.3	1,131.9
Average proportion of Genes > 50% length (%)[b]	96	86	91
Average proportion of Genes > 75% length (%)[c]	84	65	74
No. of samples with > 50% gaps	3.6	11.4	7.9
Average paralog per sample[d]	10	2.9	12.9
Length of alignment (bp)[e]	154,394	438,487	592,881
Missing data % of alignment[e]	8.3	11.7	10.2
Proportion of parsimony informative site[e]	0.41	0.46	0.44

Table 2. Summary of capture efficiency and alignment information in probe sets. [a] Mean of total number of target sequence in which coding sequence successfully assembled. [b] proportion of genes with sequences > 50% of the mean target length. [c] proportion of genes with sequences > 75% of the mean target length. [d] Mean of total number of paralogs recovered per sample. [e] total sequence after filtering.

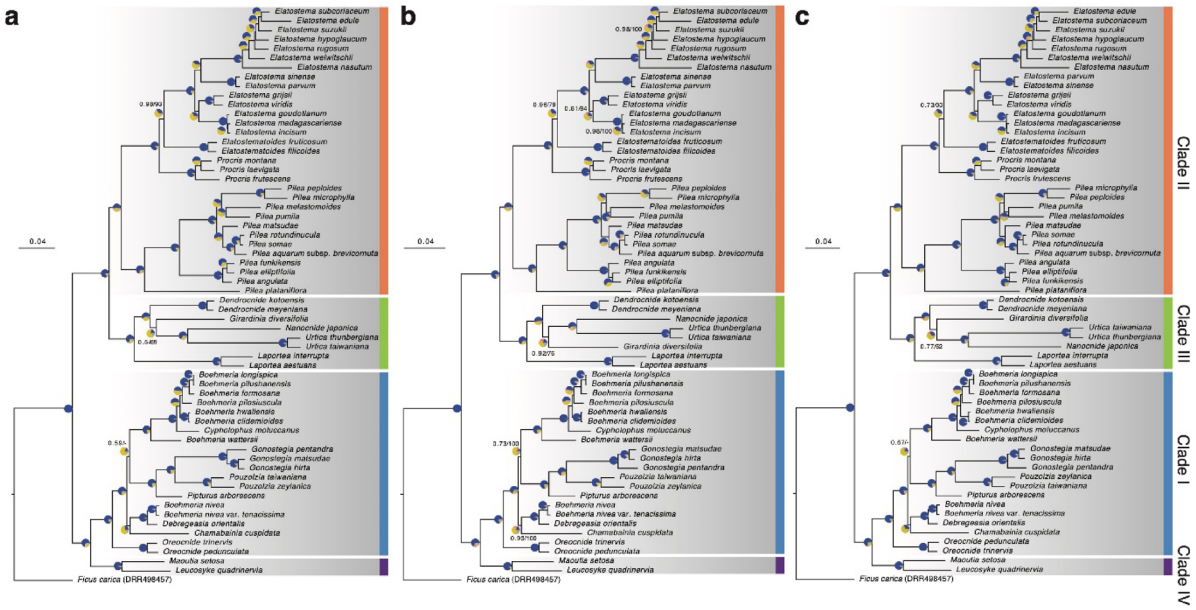

Fig. 4. Phylogenomic relationships in Urticaceae inferred from different probe sets based on RAxML analysis. The phylogenetic trees were inferred from exon sequences and reconstructed using RAxML for (**a**) Urticaceae542, (**b**) Urti-Angiosperms353, and (**c**) Urti-MarkerMiner. Numbers above each branch are the local posterior probabilities (LPP) from ASTRAL and maximum likelihood bootstrap value (BS) from RAxML. Pie charts at each node represent gene tree discordance, showing the proportion of gene trees supporting the species tree (blue), the main alternative topology (yellow), other conflicting topologies (pink), or being uninformative (grey).

Boehmeria species, including *Cypholophus moluccanus*), Clade I-2 (*Boehmeria nivea* and its varieties, *Debregeasia orientalis*, and *Chamabainia cuspidate*), and Clade I-3 (a clade containing species from *Gonostegia*, *Pipturus*, and *Pouzolzia*). ASTRAL analyses using the Urticaceae542 (exon result in Supplementary Fig. S1-a, supercontig in Supplementary Fig. S2) and Urti-MarkerMiner sets (Supplementary Fig. S1-c, S4) showed Clade I-1 and Clade I-2 as sister to Clade I-3 with low support (LPP = 0.59 & 0.67, respectively) and high gene tree discordance in PhyParts analyses. In contrast, the ASTRAL tree from Urti-Angiosperms353 set (Supplementary Fig. S1-b, S3) showed Clade I-2 and Clade I-3 as a sister group to Clade I-1 (LPP = 0.73), also exhibiting high gene tree discordance. However, all RAxML analyses strongly supported Clade I-1 and Clade I-3 as a monophyletic group, sister to Clade I-2 (all BS > 96) (Fig. 4, Supplementary Figs. S5–S7). The second conflict occurred within Clade III, involving the relationship of *Urtica*, *Nanocnide*, *Dendrocnide*, and *Giradinia*. Phylogenomic analyses using Urti-Angiosperms353 constructed by ASTRAL (Supplementary Figs. S1-b, S3), and using three bait sets constructed by RAxML (Fig. 4, Supplementary Figs. S5–S7) showed that *Dendrocnide* species was sister to the rest of species, while ASTRAL analyses using Urticaceae542 and Urti-MarkerMiner (Supplementary Figs. S1-a, S1-c, S2, S4) resolved *Giradinia* as sister to other species within Clade III. This conflict was also showed the high gene tree discordance by PhyParts analyses.

Discussion

Urticaceae542 development and comparative performance

This study presents the development and evaluation of Urticaceae542, a new Urticaceae-specific probe set. Our analysis demonstrates its effectiveness in generating robust datasets for phylogenomic analysis across clades, genera, and species. Within the Urticaceae542 set, Urti-Angiosperms353 demonstrates superior recovery efficiency, less missing data and lower sequences length variation, likely due to its foundation in the highly conserved Angiosperms353 universal probe set, which was developed from a broad sampling of over 1,100 green plant species[7,28]. Although the original Angiosperms353 target sequences were used as the reference in this study, the subsequently developed mega353 sequence by McLay et al.[29] offers improved locus recovery potential and should be evaluated in future studies. In contrast, the Urti-MarkerMiner probe set, designed with a 900 bp length threshold, yielded longer average sequences, but a higher proportion of missing data. These differences are likely due to the design of Urti-MarkerMiner probe set, which used three transcriptomes of Urticaceae based on *Arabidopsis* as a reference and focused on more divergent genomic regions, as evidenced by a greater number of parsimony-informative sites and more variable capture efficiency across genera. These findings suggest that while Urti-Angiosperms353 is well-suited for broad phylogenetic studies across clades, Urti-MarkerMiner offers better resolution for species-level relationships. This is supported by our phylogenomic analysis of *Elatostema*, where the Urti-MarkerMiner tree fully resolves all nodes within the genus, a result not achieved with Urti-Angiosperms353 alone (Fig. 4). Therefore, the combination of these two sets into Urticaceae542 provides a powerful and comprehensive tool for resolving evolutionary relationships across various taxonomic depths within the family.

Factors influencing Urticaceae542 performance

Multiple factors, such as sample age and material source, are known to influence the success of DNA retrieval and targeted gene capture[2]. Generally, fresh and silica-dried specimens typically outperform older herbarium specimens by yielding higher target enrichment efficiency and more recovered genes[30]. DNA degradation increase with sample age, also leading to a decline in target enrichment efficiency and sequence coverage[2,3]. However, our study found that neither herbarium specimens nor sample age significantly affected the performance of Urticaceae542 probe set. This finding is consistent with previous study on Urticaceae[2], which observed that in Urticaceae, herbarium specimens could even outperform silica-dried specimens in terms of enrichment efficiency for Angiosperms353 probe set. This phenomenon is likely due to the degradation of inhibitory secondary metabolites in specimens over time[2]. These compounds may interfere with DNA extraction from fresh samples. Therefore, for Urticaceae542 probe set, we conclude that sample age and material source do not appear to be critical factors influencing its performance.

While universal probe sets are designed for broad application, their capture efficiency often shows a bias toward the specific taxa used for their design[8,16,31]. Our finding with the Urticaceae542 probe set support this observation. The set was developed from transcripts of *Boehmeria nivea*, *Elatostema rivulare*, and *Urtica dioica* to represent Clades I, II, and III, respectively, within Urticaceae. As expected, we observed relatively high capture efficiency within these three genera. However, this performance advantage is not consistent observed across closely related lineages. The capture efficiency and gene recovery in Clades I, II, and III were not significantly higher than in the unsampled Clade IV. Furthermore, Clade III, which includes one of the design species (*Urtica dioica*), exhibited the lowest number of recovered loci and the shortest mean exon and supercontig lengths. These disparities are also evident at the genus level, where *Pilea* (within Clade II) showed particularly low capture efficiency, especially with the Urti-MarkerMiner probe set. These variations in probe performance likely reflect the varying genomic complexity within Urticaceae. For example, the genome of *Urtica dioica* ssp. *dioica*, exhibits highly repetitive, with over 60% of its content consisting of transposable elements and structural variation[32]. Such high repetitiveness suggests significant genomic diversity and complexity[33], which contributes to the observed variations in the capture efficiency across genera. In addition, the presence of polysaccharides and diverse secondary metabolites, which vary in composition among different Urticaceae genera[26,34], may impact DNA extraction and library preparation, further contributing to the observed differences in capture efficiency.

Pilea, the largest genus within Urticaceae with nearly 715 species[35], exhibited a notably low capture efficiency in our study. Across the 12 sampled *Pilea* species, the Urticaceae542 probe set consistently yielded shorter sequences and a lower number of recovered loci compared to other genera. This performance is not easily explained by common factors like polyploidy or ancient duplication events, as *Pilea* is a diploid genus with a moderate base chromosome number ($x = 12$)[36]. Even in *Pilea plataniflora*, which had a comparable sequence coverage (377 ± 1271) and on-target read percentages (63.2%) to other high-performing samples in this study, the recovered sequences were still short and the number of captured loci remained low, particularly with the Urti-MarkerMiner probe set. This suggests that genomic variation or sequence deletions within the genus may be hindering effective probe binding and capture. Therefore, further investigation into the genomic structure of *Pilea* is essential to fully understand the factors contributing to the low capture efficiency of the Urticaceae542 probe set.

Paralogs as indicators of polyploidy in Urticaceae

This study found significantly lower average number of paralogs per sample in the Urti-MarkerMiner probe set (2.9) compared to the Urti-Angiosperms353 set (10), which consistent with the finding that lineage-specific probe set are usually less likely to capture paralogous genes[15,37]. While overall paralog detection was low, a relatively high number of paralogs in certain species, such as *Boehmeria zollingeriana* var. *podocarpa* (48 in Urti-MarkerMiner and 94 in Urti-Angiosperms353) and *Urtica taiwaniana* (41 and 36), may suggests the occurrence of occasional polyploidy events in Urticaceae. This hypothesis is supported by documented evidence from chromosome counts and hybridization events in *Boehmeria* and *Urtica*. For example, polyploids have been

observed in *B. tricuspis* and *B. spicata*[38,39], often associated with hybridization and the formation of apomixis. Similarly, within the genus *Urtica*, two major ploidies (diploid and tetraploid) have been reported for *Urtica dioica*, along with rare triploids and pentaploids[40]. The success and widespread nature of tetraploid individuals are often attributed to allopolyploidy, which can enhance genetic diversity and fitness through effects of heterosis and the restoration of sexual reproduction following hybridization[40–42]. These findings collectively suggest that localized polyploidy, potentially driven by hybridization or apomixis, has played a role in the evolutionary history of specific lineages in Urticaceae.

Phylogenomic analysis of Urticaceae

Our phylogenomic analysis of Urticaceae, using the Urticaceae542 probe set, identified four major clades that are largely congruent with previous studies[23,27]. However, we observed some incongruences within Clade I and Clade III. Within Clade I, the placement of *Chamabainia* proved to be a source of conflict. Our analysis, which aligns with Baker et al.[27], places *Chamabainia* as the sister group to a clade containing *Debregeasia*, *Boehmeria nivea*, *Astrothalamus*, and *Archiboehmeria*. This contrasts with the phylogeny proposed by Wu et al.[23], where *Chamabainia* is not sister to these taxa. The high gene tree discordance involving placement of *Chamabainia* in our study suggests considerable phylogenetic uncertainty surrounding this genus. In addition, comparisons with phylogenetic relationships presented by Wu et al.[23] and Baker et al.[27] revealed some conflicting relationships in Clade III. For example, while *Laportea* was recovered as a monophyletic genus by Wu et al.[23], the Angiosperms353 tree from Baker et al.[27] did not support its monophyly. Due to our limited sampling of only five genera in this clade, we are unable to definitively resolve its internal classification. To address these ambiguities and achieve higher resolution, future research should incorporate more comprehensive sampling and potentially utilize Urticaceae542 probe sets.

Our results demonstrate the utility of the Urticaceae542 probe set for resolving generic and species-level relationships, particularly within taxonomically challenging genera. In this study, while most genera were recovered as monophyletic, *Boehmeria* was found to be polyphyletic, a finding congruent with its complex taxonomic history in relation to allied genera such as *Astrothalamus*, *Archiboehmeria*, *Cypholophus*, and *Debregeasia*[23,27,43]. A similar pattern of taxonomic complexity was observed in *Elatostema*, a genus with a history of controversy regarding its delimitation with allied genera such as *Elatostematoides*, *Pellionia*, and *Procris*. Tseng et al.[44] have proposed the recognition of three morphologically and phylogenetically distinct genera: *Elatostema*, *Elatostematoides*, and *Procris* and identified four well-supported clades within the redefined *Elatostema* (core *Elatostema*, *Pellionia*, *Weddellia*, and African *Elatostema*). However, phylogenetic relationships and biogeographic history within the core *Elatostema* clade have remained poorly resolved due to limited molecular data[44,45]. Our result further supports this intergeneric classification and provide stronger support for interspecific relationships. Given the documented hybrid origins[46] and the pivotal role of hybridization in driving reticulate evolution and diversification within *Elatostema*[47], comprehensive nuclear markers, like Urticaceae542, will be invaluable for evaluating hybridization events and identifying potential hybridization parents. These conclusions should be interpreted with caution, as the classical neutral theory's power to fully account for genetic diversity is debatable. Recent studies, such as those by Lake et al.[48], Lynch et al.[49], and Kreitman[50], have increasingly challenged or rejected the neutral framework. Therefore, the phylogenetic assumptions based on neutral theory may require reconsideration.

The advancements in sequencing technologies and a general decrease in sequencing costs might eventually diminish the need for target sequence capture, especially for relatively small genomes found in Urticaceae. However, the genomic data for Urticaceae remains sparse. So far, the NCBI database only lists a limited number of publicly available genome sequences for Urticaceae species. This restricted genomic availability, coupled with the difficulty of DNA extraction due to polysaccharide interference—which often compromises library quality—highlights a significant obstacle for large-scale genomic studies. Given the escalating threats to global biodiversity, our Urticaceae542 probe set, utilized in conjunction with existing natural history collections, presents a critical solution. This strategic approach not only accelerate the foundational understanding of Urticaceae, but also significantly enhances the phylogenomic potential of natural history collections.

Material & method

Urticaceae probe set design

To construct Urticaceae-specific probes from loci conserved in low copy numbers across Urticaceae, we examined transcriptome sequences of three species from Urticaceae, *Boehmeria nivea*, *Elatostema rivulare*, and *Urtica dioica*, to represent three of the four major clades (Clade I, II, III) of Urticaceae[23]. The assembly sequences of *U. dioica* (sample code: WKCY) were downloaded from the oneKP website[28,51]. The transcriptomes of *Boehmeria nivea* and *Elatostema rivulare* were sequenced in this study. Total RNAs were extracted from fresh leaves (collection information is provided in Supplementary Table S2) of these two species using the Omega E.Z.N.A. Total RNA Kit I. (Omega, Bio-tek, GA, USA) according to the manufacturer's instructions. The quantity of RNA was measured by Qubit™ 3.0 Fluorometer (Thermo Scientific, MA, USA) and quality by NanoDrop™ 2000 spectrophotometer (Thermo Scientific, MA, USA). The dual-indexed library was prepared by using Universal Plus mRNA-Seq with NuQuant (TECAN, Switzerland) and the length was evaluated by Bioanalyzer 2100 (Agilent, CA, USA) and quantities using the Qubit Fluorometer. Two libraries were sequenced by NovaSeq 6000 Sequencing System using 150 bp paired-end reads. The quality of raw reads demultiplexed by genomics core was conducted in FastQC v0.11.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmomatic v.0.36[52] with following settings: to discard reads with quality < Q20, the sliding-window set to 5:20, and then remove those cleaned reads if read length were less than 36 bp. The cleaned reads were de novo assembled using Trinity v2.15.2[53].

Subsequently, the three assembled sequences were filtered using CD-HIT[54] to remove sequences with greater than 90% identity. The putative single-copy nuclear loci were identified using MarkerMiner 1.0[10], with *Arabidopsis thaliana* as a reference and a minimum transcript length of 900 bp. For efficient sub-selection of markers for downstream probe synthesis and capture, GoldFinder[55] was employed by default settings, which included a probe coverage of 2, a total desired probe count of 30,000 and a probe length of 120 bp.

To expand our target orthologous set beyond Angiosperms353, we performed HybPiper v.2.0[56] to assemble Urticaceae-specific Angiosperms353 loci from transcriptomes of *B. nivea* and *E. rivulare*, using the general Angiosperms353 target sequence as a reference. To assess overlap between MarkerMiner-identified loci and Angiosperms353 loci, we constructed a local BLAST database from the MarkerMiner sequences using makeblastdb[57]. Each locus' sequence file from Urticaceae-specific Angiosperms353 was then queried against this database using BLASTN[57]. In cases of shared loci from MarkerMiner outcome and Angiosperms353 loci, the MarkerMiner locus was preferentially retained due to its generally greater length and higher likelihood of containing transcripts from multiple species. Subsequently, Angiosperms353 loci assembled from only one sample or with alignment lengths less than 200 bp were excluded from further analysis.

All sequences from MarkerMiner and Angiosperms353 that successfully passed aforementioned filtering steps were combined and then submitted to Arbor Biosciences (Ann Arbor, Michigan, USA) for synthesis of 80 bp long probes with 2× tiling. Before synthesis, these probes were undergoing an additional filtering step with any probes comprising ≤ 24% repeat masked, as determined by a BLAST hit to known organellar genomes.

Taxon sampling

To evaluate the capture performance of the newly developed probe set of Urticaceae, we sampled 61 species representing four major clades in Urticaceae (Supplementary Table S2). Among these, seven herbarium samples of *Elatostema* were included to assess whether sampling factors influenced capture efficiency within *Elatostema* across difference sample type (fresh vs. herbarium) and age (range of coll. year: 1855 to 2023). For phylogenomic analysis, we downloaded raw reads of *Ficus carica* (SRA: DRR498457) from the NCBI database to be an outgroup representing Moraceae, the sister family of Urticaceae.

Library preparation, hybridization and sequencing

Total DNAs were extracted from fresh and herbarium specimen following Li et al.[58]. The DNA concentration was measured by Qubit 3.0 Fluorometer (Thermo Scientific, Massachusetts, USA). The library preparation was followed by Tseng et al.[59]. The total of 500–1 μg DNA was sheared by Bioruptor Pico (Cosmo- bio Inc., Tokyo, Japan) into fragments of about 350 bp. Dual-indexed libraries were constructed using NEBNext Ultra II DNA Library Prep Kit (New England BioLabs, Massachusetts, USA), following the 300–400 bp insert size protocol.

Each of equally molar 10–12 libraries was pooled into a 2 μg pool to perform hybridisation capture of target DNA using biotinylated RNA probes from our custom-designed kit for Urticaceae (Arbor Biosciences, Ann Arbor, MI, USA). Fresh and herbarium libraries were separated into different pools. The hybridization procedure was performed at 62℃ for 20–24 h following the myBaits Custom v.5.0.3 protocol. A post-capture PCR reaction of 12 cycles were performed using KAPA HiFi HotStart ReadyMix (Kapa Biosystems, USA). Post-capture library pools were assessed for quality and quantity as before. Equimolar amount libraries of all pools were combined into one pool, and then sequenced with Illumina NovaSeq X Plus 10B using 150 bp paired-end reads at Genomics (Taipei, Taiwan).

Sequence cleaning and assembly

Raw reads were first assessed for quality using FastQC v0.11.1. (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Read trimming and filtering were performed using Trimmomatic v.0.36[52] with following settings: to discard reads with quality < Q20, the sliding-window set to 5:20, and then remove reads whose length were less than 36 bp. The cleaned reads were then assembled into contigs using HybPiper v.2.3.2[56]. This process involved mapping reads to target sequences with BWA v.0.7.19[60], followed by assembly with SPAdes v.3.13.1[61]. In addition to the default HybPiper coding sequence (CDS, exon sequence) output extracted with exonerate v.2.2[62], the optional HybPiper intronerate.py script was run to also extract supercontig sequences, which contain both CDS and non-coding flanking sequence. Finally, potential paralogs, as indicated by the presence of multiple long contigs for a locus, were identified using HybPiper scripts.

Phylogenomic analysis

The exon and supercontig sequences were both conducted for the phylogenomic analyses. To mitigate potential phylogenetic error, we excluded loci with less than 80% samples recovery and with paralog warning in more than five species for further phylogenomic analyses. Every remaining locus was aligned using the automatic alignment strategy in MAFFT v.7.525[63], followed by trimming with trimAl v.1.5.0 using automated1 setting[64]. For the concatenated-based analysis, all alignments were combined with AMAS[65] and then Maximum Likelihood (ML) tree was reconstructed using RAxML v.8.2.12[66] with the GTRGAMMA model with 100 searches for the best tree and 100 standard bootstrap (BS) replicates. For the coalescent-based analysis, individual gene trees were reconstructed using RAxML v.8.2.12 under the same parameters. To improve the accuracy of species tree inference, we first collapsed branches with less than 10% bootstrap support before using ASTRAL v.5.7.8[67] to summarize the gene trees into a final species tree estimate. Quartet support for ASTRAL species trees was evaluated using local posterior probabilities (LPP)[68]. We also conducted a concordance analysis using PhyParts[69] with a bootstrap threshold of 33%. The resulting output was used to quantify the numbers of supporting and conflicting bipartitions, which were subsequently visualized as pie charts on the species tree and RAxML concatenated tree based on exon sequences.

Statistics

To evaluate if sequencing and assembly metrics differed significantly among four major clades (Clade I–IV) established by Wu et al.[23] and genera, we conducted a series of statistical test. We first assessed the normality of each metric, including number of reads, sequence length, supercontig length, on-target reads, and number of genes recovered, within each clade and genus using the Shapiro–Wilk test[70]. For metrics that were normally distributed across all dataset ($p > 0.01$), we applied one-way ANOVA followed by Tukey's HSD test for pairwise comparisons. For metrics that violated the normality, we used the non-parametric Kruskal–Wallis test[71]. We applied the same procedure to evaluate differences between sample types (herbarium vs. fresh material) from *Elatostem* samples. Given the smaller sample size and the two-group comparison, we additionally used Mann–Whitney U test[72] test to provide robust inference across metrics.

To examine differences in variance across genera within each probe set (designed from MarkerMiner and Angiosperms353), we conducted Levene's test[73]. Subsequently, we used paired *t*-tests to compare the performance of the two probe sets, focusing on key metrics such as recovery efficiency, sequence length, missing data, and phylogenetic informativeness. A Pearson correlation analysis was employed to assess the relationship between percentage of on-target reads and sequence length, and sample collection year of *Elatostema* affecting sequencing and capture performance. All statistical analyses were performed in R version 4.3.2, using the car[74] and stats packages[75].

Data availability

The raw sequencing data from this study have been submitted to NCBI BioProject (http://www.ncbi.nlm.nih.gov/bioproject) under BioProject number PRJNA1332059. The datasets generated and analyzed during the current study, including probe sequences, sequence assemblies, alignments, and phylogenetic trees, are available in the GitHub repository: https://github.com/yuhsintseng/Urticaceae542.

Received: 18 August 2025; Accepted: 3 December 2025
Published online: 15 December 2025

References

1. McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D. & Yang, Y. Practical considerations for plant phylogenomics. *Appl. Plant. Sci.* **6**, e1038. https://doi.org/10.1002/aps3.1038 (2018).
2. Brewer, G. E. et al. Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Front. Plant. Sci.* **10**, 1102. https://doi.org/10.3389/fpls.2019.01102 (2019).
3. Hart, M. L., Forrest, L. L., Nicholls, J. A. & Kidner, C. A. Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* **65**, 1081–1092. https://doi.org/10.12705/655.9 (2016).
4. Couvreur, T. L. P. et al. Phylogenomics of the major tropical plant family Annonaceae using targeted enrichment of nuclear genes. *Front. Plant Sci.* 9, (1941). https://doi.org/10.3389/fpls.2018.01941 (2019).
5. Mandel, J. R. et al. A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the compositae. *Appl. Plant. Sci.* **2**, 1300085. https://doi.org/10.3732/apps.1300085 (2014).
6. Slimp, M., Williams, L. D., Hale, H. & Johnson, M. G. On the potential of Angiosperms353 for population genomic studies. *Appl. Plant. Sci.* **9**, e11419. https://doi.org/10.1002/aps3.11419 (2021).
7. Johnson, M. G. et al. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Syst. Biol.* **68**, 594–606. https://doi.org/10.1093/sysbio/syy086 (2019).
8. Yardeni, G. et al. Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. *Mol. Ecol. Resour.* **22**, 927–945. https://doi.org/10.1111/1755-0998.13523 (2022).
9. Medina, N., Tank, D. C. & Espindola, A. Calceolariaceae809: A bait set for targeted sequencing of nuclear loci. *Appl. Plant. Sci.* **11**, e11557. https://doi.org/10.1002/aps3.11557 (2023).
10. Chamala, S. et al. MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Appl. Plant. Sci.* **3**, 1400115. https://doi.org/10.3732/apps.1400115. (2015).
11. Lautenschlager, U., Scheunert, A. & LoCoLotive In Silico mining for low-copy nuclear loci based on target capture probe sets and arbitrary reference genomes. *Appl. Plant. Sci.* **11**, e11535. https://doi.org/10.1002/aps3.11535 (2023).
12. McDonnell, A. J. et al. Exploring Angiosperms353: developing and applying a universal toolkit for flowering plant phylogenomics. *Appl. Plant. Sci.* **9**, e11443. https://doi.org/10.1002/aps3.11443 (2021).
13. Hendriks, K. P. et al. The best of both worlds: Combining lineage-specific and universal bait sets in target-enrichment hybridization reactions. *Appl. Plant. Sci.* **9**, e11438. https://doi.org/10.1002/aps3.11438 (2021).
14. Wang, X. Q., Xiong, T., Wang, Y. Y., Zhang, X. R. & Sun, M. Integrating genomic sequencing resources: an innovative perspective on recycling with universal Angiosperms353 probe sets. *Hortic. Adv.* **2**, 4. https://doi.org/10.1007/s44281-023-00026-z (2024).
15. Jones, K. E. et al. An empirical assessment of a single family-wide hybrid capture locus set at multiple evolutionary timescales in Asteraceae. *Appl. Plant. Sci.* **7**, e11295. https://doi.org/10.1002/aps3.11295 (2019).
16. Chau, J. H., Rahfeldt, W. A. & Olmstead, R. G. Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. *Appl. Plant. Sci.* **6**, e1032. https://doi.org/10.1002/aps3.1032 (2018).
17. Eserman, L. A., Thomas, S. K., Coffey, E. E. D. & Leebens-Mack, J. H. Target sequence capture in orchids: developing a kit to sequence hundreds of single-copy loci. *Appl. Plant. Sci.* **9**, e11416. https://doi.org/10.1002/aps3.11416 (2021).
18. Gomez, M. S. et al. A customized nuclear target enrichment approach for developing a phylogenomic baseline for *Dioscorea* Yams (Dioscoreaceae). *Appl. Plant. Sci.* **7**, e11254. https://doi.org/10.1002/aps3.11254 (2019).
19. Pezzini, F. F. et al. Target capture and genome skimming for plant diversity studies. *Appl. Plant. Sci.* **11**, e11537. https://doi.org/10.1002/aps3.11537 (2023).
20. Forrest, L. L. et al. The limits of Hyb-Seq for herbarium specimens: impact of preservation techniques. *Front. Ecol. Evol.* **7**, 439. https://doi.org/10.3389/fevo.2019.00439 (2019).
21. POWO. Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Published on the Internet. Nov (2025). https://powo.science.kew.org/. Accessed 1.
22. Hadiah, J. T., Conn, B. J. & Quinn, C. J. Infra-familial phylogeny of Urticaceae, using Chloroplast sequence data. *Aust Syst. Bot.* **21**, 375–385. https://doi.org/10.1071/SB08041 (2008).
23. Wu, Z. Y. et al. Molecular phylogeny of the nettle family (Urticaceae) inferred from multiple loci of three genomes and extensive generic sampling. *Mol. Phylogenet Evol.* **69**, 814–827. https://doi.org/10.1016/J.Ympev.2013.06.022 (2013).
24. Wu, Z. Y. et al. Testing darwin's transoceanic dispersal hypothesis for the inland nettle family (Urticaceae). *Ecol. Lett.* **21**, 1515–1529. https://doi.org/10.1111/ele.13132 (2018).

25. Wu, Z. Y. et al. Ancestral state reconstruction reveals rampant homoplasy of diagnostic morphological characters in Urticaceae, conflicting with current classification schemes. *PLoS One*. **10**, e0141821. https://doi.org/10.1371/journal.pone.0141821 (2015).
26. Sarrazola, J., Alzate, F. & Obtaining DNA from urticaceae: Overcoming the challenges associated with chemical compounds and herbarium specimens. *Int. J. Mol. Biol.* **4**, 158–165. https://doi.org/10.15406/ijmboa.2019.04.00113 (2019).
27. Baker, W. J. et al. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Syst. Biol.* **71**, 301–319. https://doi.org/10.1093/sysbio/syab035 (2022).
28. Leebens-Mack, J. H. et al. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685. https://doi.org/10.1038/s41586-019-1693-2 (2019).
29. McLay, T. G. B. et al. New targets acquired: Improving locus recovery from the Angiosperms353 probe set. *Appl. Plant. Sci.* **9**, e1142. https://doi.org/10.1002/aps3.11420 (2021).
30. Villaverde, T. et al. Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. *New. Phytol.* **220**, 636–650. https://doi.org/10.1111/nph.15312 (2018).
31. Bentz, P. C. & Leebens-Mack, J. D. Asparagaceae1726: an Asparagaceae-specific probe set targeting 1726 loci for Hyb-Seq and phylogenomics in the family. *Appl. Plant. Sci.* **12**, e11597. https://doi.org/10.1002/aps3.11597 (2024).
32. Hirabayashi, K. et al. A high-quality phased genome assembly of stinging nettle (Urtica dioica ssp. dioica). *Plants-Basel* **14**, 124. https://doi.org/10.3390/plants14010124 (2025).
33. Lee, S. & Kim, N. S. Transposable elements and genome size variations in plants. *Genomics Inf.* **12**, 87–97. https://doi.org/10.5808/GI.2014.12.3.87 (2014).
34. Gonçalves, F. M. B., Sousa, A. L. & Oliveira, R. R. Use of mass spectrometry as a tool for the search or identification of flavonoids in Urticaceae. *Rodriguésia* 74, e01152022. (2023). https://doi.org/10.1590/2175-7860202374049
35. Monro, A. K. Revision of species-rich genera: A phylogenetic framework for the strategic revision of *Pilea* (Urticaceae) based on cpDNA, nrDNA, and morphology. *Am. J. Bot.* **93**, 426–441. https://doi.org/10.3732/ajb.93.3.426 (2006).
36. Ferdousee, N., Denda, T. & Yokota, M. Karyological observations on Dwarf individuals of *Pilea brevicornuta* Hayata (Urticaceae) growing in the riparian habitat in Amami-Oshima of the Ryukyus, Japan. *Chromosome Sci.* **18**, 3–8. https://doi.org/10.11352/scr.18.3 (2015).
37. Jantzen, J. R. et al. A two-tier bioinformatic pipeline to develop probes for target capture of nuclear loci with applications in melastomataceae. *Appl. Plant. Sci.* **8**, e11345. https://doi.org/10.1002/aps3.11345 (2020).
38. Yahara, T. Evolution of agamospermous races in *Boehmeria* and *Eupatorium*. *Plant. Spec. Biol.* **5**, 183–196. https://doi.org/10.1111/j.1442-1984.1990.tb00203.x (1990).
39. Tang, Q. et al. A full-length reference floral transcriptome of Boehmeria tricuspis provides insights into apomeiosis and polyploidy. *Int. J. Genomics*. **2019** (4025747). https://doi.org/10.1155/2019/4025747 (2019).
40. Rejlová, L. et al. Polyploid evolution: the ultimate way to Grasp the nettle. *PLoS One*. **14**, e0218389. https://doi.org/10.1371/journal.pone.0218389 (2019).
41. te Beest, M. et al. The more the better? The role of polyploidy in facilitating plant invasions. *Ann. Bot.* **109**, 19–45. https://doi.org/10.1093/aob/mcr277 (2012).
42. Rejlová, L. et al. Disparity between morphology and genetics in *Urtica dioica* (Urticaceae). *Bot. J. Linn. Soc.* **195**, 606–621. https://doi.org/10.1093/botlinnean/boaa076 (2021).
43. Zhan, M. et al. Polyphyly of *Boehmeria* (Urticaceae) congruent with plastome structural variation. *Front. Plant. Sci.* **15**, 1297499. https://doi.org/10.3389/fpls.2024.1297499 (2024).
44. Tseng, Y. H., Monro, A. K., Wei, Y. G. & Hu, J. M. Molecular phylogeny and morphology of *Elatostema* s. l. (Urticaceae): implications for inter- and infrageneric classifications. *Mol. Phylogenet Evol.* **132**, 251–264. https://doi.org/10.1016/j.ympev.2018.11.016 (2019).
45. Tseng, Y. H., Monro, A. K. & Hu, J. M. Biogeographic patterns and evolutionary history of *Elatostema* (Urticaceae). *Bot. Stud.* **66**, 10. https://doi.org/10.1186/s40529-025-00456-0 (2025).
46. Tseng, Y. H. & Hu, J. M. A new hybrid from Taiwan, *Elatostema xhybrida* (Urticaceae), is the first confirmed natural hybrid for urticaceae. *Phytotaxa* **161**, 43–60. https://doi.org/10.11646/phytotaxa.161.1.2 (2014).
47. Fu, L. F. et al. Heteroploidy hybridization and apomixis facilitate the origin of a new *Elatostema* species: insights from evidence of phylogenomics, flow cytometry, cytology and morphology. *J. Syst. Evol.* https://doi.org/10.1111/jse.13199 (2025).
48. Lake, N. J. et al. Quantifying constraint in the human mitochondrial genome. *Nature* **635**, 390–397. https://doi.org/10.1038/s41586-024-08048-x (2024).
49. Lynch, M., Wei, W., Ye, Z. Q. & Pfrender, M. The genome-wide signature of short-term temporal selection. *Proc. Natl. Acad. Sci. U.S.A.* 121, e2307107121. (2024). https://doi.org/10.1073/pnas.2307107121
50. Kreitman, M. The neutral theory is dead. Long live the neutral theory. *Bioessays* **18**, 678–683. https://doi.org/10.1002/bies.950180812 (1996).
51. Carpenter, E. J. et al. Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). *Gigascience* 8, giz126. (2019). https://doi.org/10.1093/gigascience/giz126
52. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* **30**, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 (2014).
53. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652. https://doi.org/10.1038/nbt.1883 (2011).
54. Fu, L. M., Niu, B. F., Zhu, Z. W., Wu, S. T. & Li, W. Z. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565 (2012).
55. Vargas, O. M., Heuertz, M., Smith, S. A. & Dick, C. W. Target sequence capture in the Brazil nut family (Lecythidaceae): marker selection and capture from genome skimming data. *Mol. Phylogenet Evol.* **135**, 98–104. https://doi.org/10.1016/j.ympev.2019.02.020 (2019).
56. Johnson, M. G. et al. Hybpiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant. Sci.* **4**, 1600016. https://doi.org/10.3732/apps.1600016 (2016).
57. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421. https://doi.org/10.1186/1471-2105-10-421 (2009).
58. Li, D. K., Huang, C. L., Tian, J. B., Wang, Y. K. & Wang, Y. Q. Extraction ways of high qualitiful DNA from *Z. jujuba* mill. *Mol. Plant. Breed.* **3**, 579–583. https://doi.org/10.1007/978-1-0716-2517-0_5 (2005).
59. Tseng, Y. H., Chien, H. C. & Zhu, G. X. Comparative plastome analyses and phylogenetic insights of *Elatostema*. *BMC Plant. Biol.* **25**, 537. https://doi.org/10.1186/s12870-025-06569-9 (2025).
60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595. https://doi.org/10.1093/bioinformatics/btp324 (2010).
61. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477. https://doi.org/10.1089/cmb.2012.0021 (2012).
62. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* **6**, 13. https://doi.org/10.1186/1471-2105-6-31 (2005).
63. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. https://doi.org/10.1093/molbev/mst010 (2013).
64. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348 (2009).

65. Borowiec, M. L. AMAS: a fast tool for alignment manipulation and computing of summary statistics. *Peer J.* **4**, 1660. https://doi.org/10.7717/peerj.1660 (2016).
66. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 (2014).
67. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* **19**, 153. https://doi.org/10.1186/s12859-018-2129-y (2018).
68. Sayyari, E. & Mirarab, S. Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* **33**, 1654–1668. https://doi.org/10.1093/molbev/msw079 (2016).
69. Smith, S. A., Moore, M. J., Brown, J. W. & Yang, Y. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* **15**, 150. https://doi.org/10.1186/s12862-015-0423-0 (2015).
70. Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591–611. https://doi.org/10.1093/biomet/52.3-4.591 (1965).
71. Kruskal, W. H. & Wallis, W. A. Use of ranks in one–criterion variance analysis. *J. Am. Stat. Assoc.* **47**, 583–621. https://doi.org/10.2307/2280779 (1952).
72. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. Ann. *Mat. Stat.* **18**, 50–60 (1947).
73. Levene, H. in *In Contributions To Probability and Statistics*. 278–292 (eds Olkin, I.) (Stanford University Press, 1960).
74. Fox, J. & Weisberg, S. *An R Companion to Applied Regression, Third edition*, (2019). https://www.john-fox.ca/Companion/
75. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, (2022).

Acknowledgements
We are grateful to Chia-Lun Hsieh and Geng-Xi Zhu for assisting with the fieldwork. The study was supported by the National Science and Technology Council, Taiwan, to YHT (MOST 111-2621-B-005-001-MY3).

Author contributions
YHT conceptualized the study, collected the samples, conducted the experiments, and drafted the manuscript. HCC conducted the experiments and drafted the manuscript. All authors read and approved the final manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-31584-z.

Correspondence and requests for materials should be addressed to Y.-H.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.