scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A DNA barcode reference of Asian ferns with expert-identified voucher specimens and DNA samples

Li-Yaung Kuo[1]✉, Sheng-Kai Tang[1], Yu-Hsuan Huang[1], Pei-Jun Xie[1,19], Cheng-Wei Chen[2,3,4], Zhi-Xiang Chang[5], Tian-Chuan Hsu[6], Yi-Han Chang[6], Yi-Shan Chao[4], Chien-Wen Chen[6], Susan Fawcett[7,8], Joel H. Nitta[9], Michael Sundue[10], Tzu-Tong Kao[11,12], Hong Truong Luu[13], Andi Maryani A. Mustapeng[14,15], Fulgent P. Coritico[16,17], Victor B. Amoroso[16,17] & Yong Kien Thai[18]

Ferns belong to species-rich group of land plants, encompassing more than 11,000 extant species, and are crucial for reflecting terrestrial ecosystem changes. However, our understanding of their biodiversity hotspots, particularly in Southeast Asia, remains limited due to scarce genetic data. Despite harboring around one-third of the world's fern species, less than 6% of Southeast Asian ferns have been DNA-sequenced. In this study, we addressed this gap by sequencing 1,496 voucher-referenced and expert-identified fern samples from (sub)tropical Asia, spanning Malaysia, the Philippines, Taiwan, and Vietnam, to retrieve their *rbcL* and *trnL-F* sequences. This DNA barcode collection of Asian ferns encompasses 956 species across 152 genera and 34 families, filling major gaps in fern biodiversity understanding and advancing research in systematics, phylogenetics, ecology and conservation. This dataset significantly expands the Fern Tree of Life to over 6,000 species, serving as a pivotal and global reference for worldwide barcoding identification of ferns.

Background & Summary

A diverse modern land plant lineages, ferns are estimated to include more than 11,000 species[1]. These plants are abundant and most diverse in tropical and insular regions in the world[2]. Our understanding of the hyper-diversity in these areas began during floristic investigations of the 19th and 20th centuries, before molecular techniques became available. However, tropical ferns from these diversity hotspots remain scarce in DNA databases. This is especially true for Southeast Asia, where around one-third of world fern species are concentrated[3], but less than 6% of species have been sequenced[4].

[1]Institute of Molecular and Cellular Biology, National Tsing Hua University, Hsinchu City, 30013, Taiwan. [2]Biodiversity Program, Taiwan International Graduate Program, Academia Sinica and National Taiwan Normal University, Taipei, 115, Taiwan. [3]Biodiversity Research Center, Academia Sinica, Taipei, 115, Taiwan. [4]Department of Life Science, National Taiwan Normal University, Taipei, 106, Taiwan. [5]Fushan Research Center, Taiwan Forestry Research Institute, Yilan, 264013, Taiwan. [6]Taiwan Forestry Research Institute, Taipei, 10066, Taiwan. [7]National Tropical Botanical Garden, Kalaheo, Kauai, Hawaii, USA. [8]University and Jepson Herbaria, University of California, Berkeley, California, USA. [9]Graduate School of Global and Transdisciplinary Studies, College of Liberal Arts and Sciences, Chiba University, Chiba, 263-8522, Japan. [10]The Royal Botanic Garden Edinburgh, 20a Inverleith Row, Edinburgh, UK. [11]Department of Biological Sciences, National Sun Yat-sen University, Kaohsiung, 804201, Taiwan. [12]Institute of Plant and Microbial Biology, Academia Sinica, Taipei, 115201, Taiwan. [13]Southern Institute of Ecology, Institute of Applied Materials Science, Vietnam Academy of Science and Technology, Ho Chi Minh City, Vietnam. [14]Forest Research Centre, Sabah Forestry Department, 90715, Sandakan, Sabah, Malaysia. [15]Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah, 88400, Kota Kinabalu, Sabah, Malaysia. [16]Plant Biology Division, Institute of Biological Sciences, College of Arts and Sciences, Central Mindanao University, University Town, Musuan, Bukidnon, 8710, the Philippines. [17]Center for Biodiversity Research and Extension in Mindanao (CEBREM), Central Mindanao University, University Town, Musuan, Bukidnon, 8710, the Philippines. [18]Institute of Biological Sciences, Faculty of Science, University Malaya, 50603, Kuala Lumpur, Malaysia. [19]Present address: Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, New York, USA. ✉e-mail: lykuo@life.nthu.edu.tw

DNA barcoding — sequencing DNA regions demonstrated to be of broad taxonomic utility — has proven to be an effective tool to evaluate genetic diversity from taxon-wide collections[5]. However, choosing standard loci for DNA barcoding is critical to compare taxa and samples with diverse origins. Like many other plants, plastid *rbcL* (ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit) and *trnL-F* (the intergenic spacer between tRNA-Leu and tRNA-Phe genes; sometimes referred to the region extending to the tRNA-Leu intron) are commonly used DNA barcodes in ferns because they have universal primers[6–8], and, as plastid DNA regions, they are uniparentally inherited (reviewed in Kuo *et al.*[9]). DNA-referencing of these two barcodes has been widely employed[6,10–14] and highly successful in fern phylogenetic studies[15]; to date, more than 5,600 fern species are available with at least one of the two barcode sequences in GenBank[15]. *trnL-F* has been shown to have greater interspecific and intraspecific variation relative to *rbcL* because the former is include (a) non-coding spacer(s), whereas the latter is a protein-coding gene. *trnL-F* has been previously shown in studies including smaller sampling of ferns to have higher species discrimination rates than *rbcL*[8,16]. In comparison, *rbcL* is useful as a phylogenetic marker at deeper divergence levels due to its slower evolution rate, and is the most-frequently sequenced genetic region in ferns[15] as well as a core DNA barcode in land plants[17]. Therefore, sequencing both regions is highly recommended for fern DNA barcoding projects, which can serve to both identify species and expand the phylogenetic sampling of the fern tree of life. *trnH-psbA*, another frequently used non-coding DNA barcode in plants[17], is not a prevalent choice for ferns due to its relatively slow substitution rate in most species[8,18]. Other proposed plant DNA barcodes, such as *nrITS* and *matK*, lack (one of) the above-mentioned advantages, and are thus not prioritized in fern DNA barcoding studies[8,19].

DNA barcoding has been successfully applied in various ecological and floristic surveys of ferns, and is particularly useful for DNA-identification of their cryptic gametophyte stage[7,13,16,20,21]. Fern gametophytes are free-living but frequently have too few morphological features to be reliably determined to species[16]. With DNA approaches, phenological studies and habitat investigations of fern gametophytes in the field can be accomplished based on reliably identified samples (e.g. Quinlan *et al.*[22] and Wu *et al.*[7]). Notably, some ferns have populations consisting of long-lived gametophytes but producing no spore-producing individuals, which are referred to as 'independent gametophytes'[23–25]. To confirm their species identities, such a molecular identification tool is indispensable. Additionally, fern DNA barcodes have been used to study novel ecological links between these plants and other organisms, including insect pollinators[26] and rhizobium bacteria[27]. As demonstrated in earlier research, prolific production coupled with high dispersibility means that fern spores provide a key signature reflecting environmental changes[28,29]. The emerging field of environmental DNA research relies on further development and publication of DNA barcodes, enabling ecologists to readily monitor environmental dynamics and to expand documentation of biodiversity[30]. A comprehensive and global database of fern DNA barcodes is therefore essential.

Here, we present a voucher-referenced and expert-identified collection of Asian ferns with DNA barcode regions *rbcL* and *trnL-F*, encompassing 1,496 samples from 956 species, including hybrid taxa. Of particular value is the large proportion of samples from fern diversity hotspots in South-eastern Asia, including the Philippines, Vietnam, and Malaysia, which fill major gaps in our understanding of these plants, and will facilitate future research of understanding fern diversity there. Furthermore, this DNA barcode dataset also serves as valuable resources for advancing investigation in systematics, phylogenetics, and conservation genetics of ferns from these biodiversity hotspots. Among these samples, 292 species were sequenced for the first time with these DNA barcodes, and will contribute to a notable expansion (4.6%) to the Fern Tree of Life (FTOL)[15]. The incorporation of our new sequence dataset with those already existing in FTOL offers the pivotal global database for fern barcode identification.

Methods

Sampling and specimen identification. We sampled a total of 1,496 fern collections across 390 localities in Malaysia, the Philippines, Taiwan, and Vietnam. They were collected during field expeditions spanning from 2005 to 2022, and were vouchered with specimens in Taiwan Forestry Research Institute Herbarium (TAIF). From each collection, tissue was preserved on silica gel for DNA extractions, which are also publicly available at the TPG website (https://www.twfern.org/DB/DNACollection). Species-level identifications were conducted by experienced fern taxonomists, relying on the morphology and genetic data of voucher specimens. A few collections may represent undescribed species or require further taxonomic investigation, so their identification was thus determined only to the generic level.

DNA extraction, amplicon preparation, and sequencing. The workflow of this study is summarized in Fig. 1. First, fresh or silica-dried leaf tissues were used for DNA extractions of the 1,496 fern samples following the CTAB protocol by Kuo[31]. To amplify these DNA barcodes, various PCRs of *rbcL* and *trnL-F* amplicons were carried out according to three different sequencing methods, including (1) the traditional Sanger sequencing, and the multiplexing strategies utilizing high-throughput next-generation sequencing (NGS), (2) PacBio CCS (circular consensus sequencing) and (3) Illumina MiSeq. For Sanger and PacBio CCS, longer amplicons (~1 kbp) were amplified and sequenced, with *trnL* intron alongside *trnL-F*. For Illumina MiSeq, shorter amplicons (<600 bp) were amplified and sequenced, and the *rbcL* region was split into two amplicons, *rbcL*N and *rbcL*C (Supplementary Figure 1). Dual 8nt-indexed primer sets were employed for the multiplexed amplicons for PacBio CCS and Illumina MiSeq. For these primers, the conservative priming regions were same as Wu *et al.*[7] or identified using their approach, with different 8nt-indexes added to the 5′ ends. In addition, for Illumina MiSeq, we co-amplified *trnL-F* and *rbcL*N in the same PCR reaction for each of individual DNA sample. The details of primer sets and thermal conditions of PCR cycles were provided in Supplementary Tables 1 and 2, respectively. Each PCR reaction comprised 7.5 μL 2 × SuperRed PCR Master Mix (BIOTOOLS Co., Ltd., New Taipei, Taiwan), 2 μL DNA template (10 ng/μL), 0.75 μL of each primer (10 nM), and ddH2O added to a total volume of 15 μL.

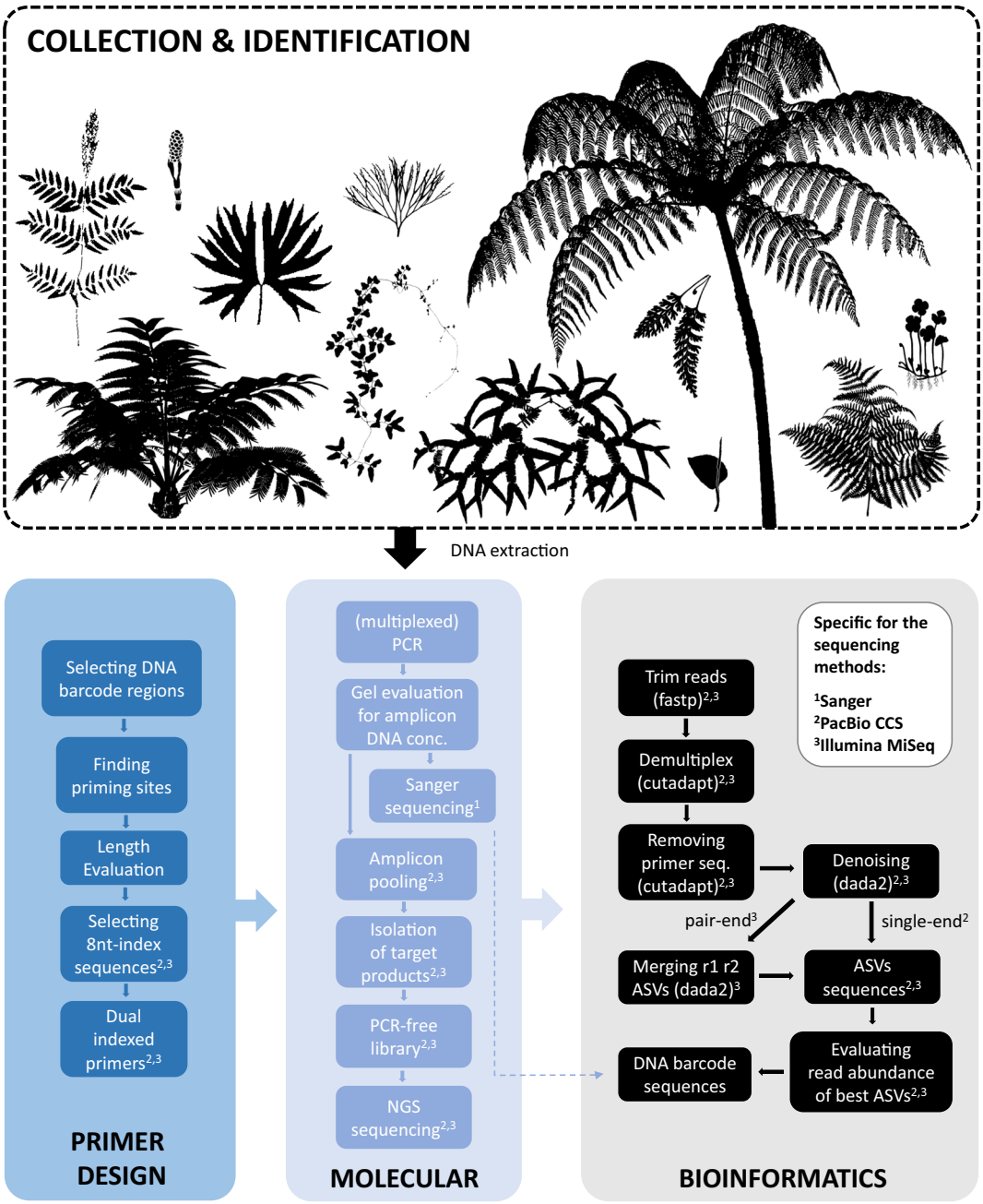

Fig. 1 The workflow of collecting fern DNA barcodes. The fern diagrams were downloaded from https://www.phylopic.org/ or modified from Vasco *et al.*[47].

For Sanger sequencing, PCR products were first purified using ExoSAP-IT (Thermo Fisher Scientific, Waltham, Massachusetts, USA), and then sequenced by an ABI 3730XL DNA Analyzer (Thermo Fisher Scientific, Waltham, Massachusetts, USA). For PacBio CCS and Illumina MiSeq, amplicon products were initially assessed through 1 × TAE 1% agarose gels. We then pooled these products to achieve similar molecular concentrations according to their estimated DNA concentrations. The DNA fragments with target sizes in these multiplexed amplicon pools were isolated by electrophoreses with 1 × TAE 0.8% agarose gels and purified using Geneaid Large DNA Fragments Extraction Kit (Geneaid, New Taipei, Taiwan). For the isolated DNA products with low O.D. values (i.e. A260/280 and A260/230 < 1.8), further purification was conducted using 1 × AMPure XP Beads (Beckman Coulter, Brea, California, USA) before NGS library constructions. PacBio CCS library preparation and sequencing were carried out at the Sequencing and Genomic Technologies Core Facility of the Duke University Center for Genomic and Computational Biology, utilizing a single SMRT cell on a PacBio sequel sequencer with 3.0 chemistry (Pacific Biosciences, Menlo Park, California, USA). The fastq reads of PacBio CCS were then employed for the downstream demultiplexing and ASV (amplicon sequence variant) generation (see below). We constructed PCR-free libraries for Illumina MiSeq using KAPA Dual-Index Adapter Kit (Roche, Basel, Switzerland). DNA molecular concentrations of these Illumina libraries were measured

	rbcL	*trnL-F*
Sanger sequencing	305	379
PacBio CCS	239	241
Illumina MiSeq	948	720

Table 1. Number of DNA barcode sequences by different sequencing methods.

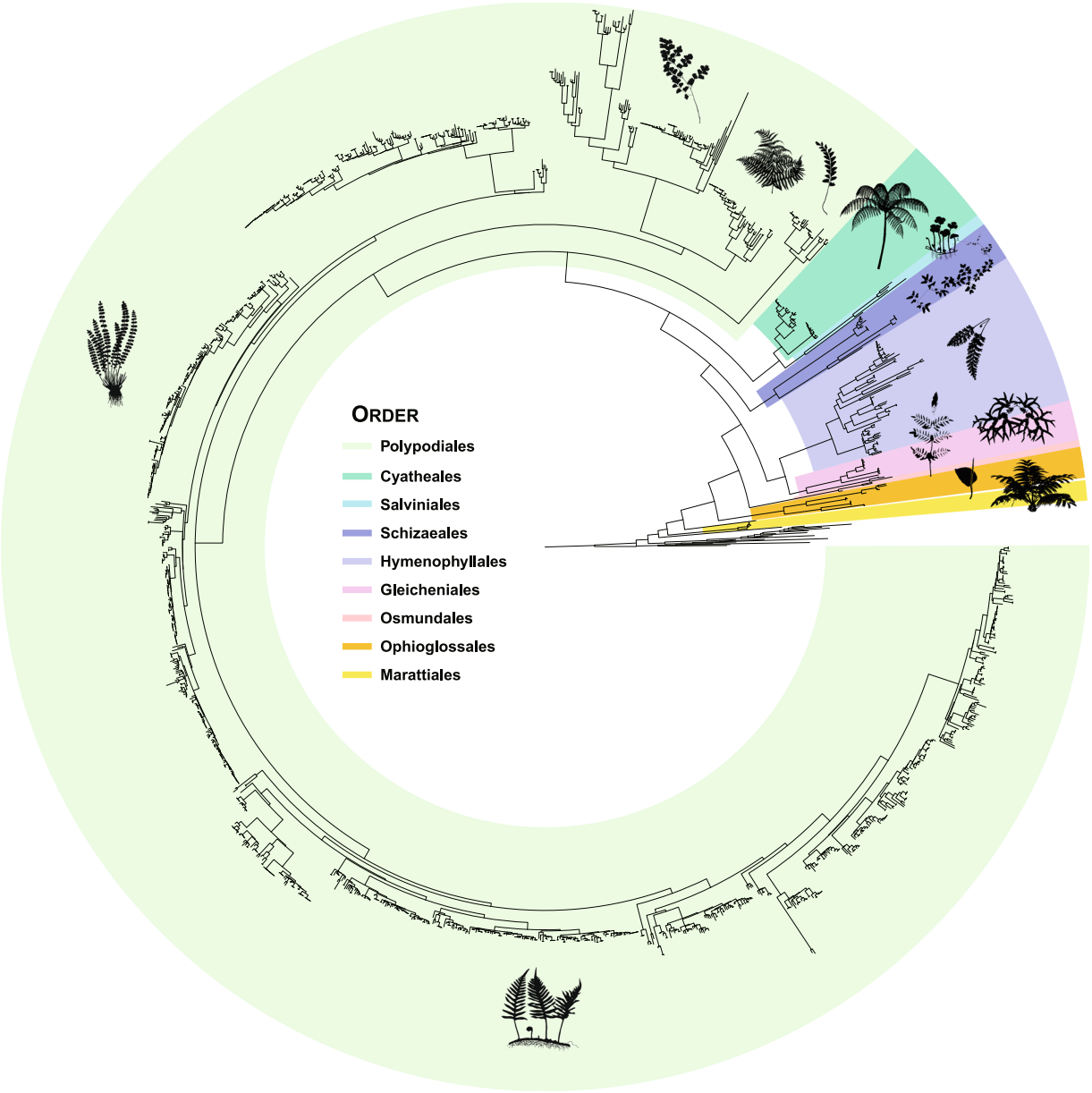

Fig. 2 Maximum-likelihood *rbcL* + *trnL-F* phylogeny noted with the order-level taxonomy sensu PPG I. For details about each test, see Methods and Technical Validation. The fern diagrams were downloaded from https://www.phylopic.org/ or modified from Vasco *et al.*[47] and Dong *et al.*[48]. The Psilotales and Equisetales are not colored.

using Sequencing Library qPCR Quantification (Illumina, San Diego, California, USA). The libraries were then sequenced on Illumina's MiSeq PE300 platform using Reagent Kit v3 (600-cycle; Illumina, San Diego, California, USA). To obtain DNA barcode sequences from the NGS fastq reads, adapter sequences were first trimmed by fastp[32]. Cutadapt[33] was then applied for demultiplexing and removal of primer sequences, and dada2[34] was finally used to generate their ASV sequences. The most abundant ASV from each sample was selected as the DNA barcode sequence for further analyses.

Data verification. For DNA barcodes obtained through the NGS strategies we evaluated the read abundance and proportion of the best ASV sequence per sample. Sequences with abundances below 30 reads for Illumina MiSeq and 10 reads for PacBio CCS were excluded from further analysis due to potential contamination. For Illumina MiSeq, we considered the process of index hopping, where indexed pair-end reads from different samples could contaminate each other. Additionally, samples with the best ASV sequences accounting for proportions lower than 0.9 were inspected to identify potential contamination by nonspecific PCR products or other DNA sources. By these, some *rbcL* amplicons were found likely to contain plastid-derived copies from mitochondrial genomes. In such cases, we manually retrieved sequences from the original ASVs and found the correct copy. Finally, for both *rbcL* and *trnL-F* barcodes, we used MUSCLE[35] to align all DNA sequences, and FastTree2[36] to reconstruct a preliminary phylogeny. These analyses aimed to identify samples that were potentially misidentified or mislabelled. For the formal phylogeny, we first aligned all specimens within each family using MAFFT[37]. We merged these family-level *trnL-F* alignments into a single alignment using the MAFFT–merge argument, and aligned *rbcL* with the same outgroups as those used in Nitta *et al.*[15]. We then inferred maximum-likelihood (ML) phylogenies each with 1000 ultrafast bootstrap replicates (UFBS)[38] using IQTREE[39]

Data Records

In total, we included 1,492 *rbcL* and 1,340 *trnL-F* DNA barcode sequences (Table 1) from 956 identified species across 152 genera, 34 families, and 11 orders. Among them, 22 and 23 species/taxa belong to hybrids and species complexes, respectively (Supplementary Tables 3 and 4). Except for 21 *rbcL* and 12 *trnL-F* sequences published earlier (Supplementary Table 5), all are newly generated in this study. Sequence information including their voucher, GenBank accession numbers, and sequencing methods are provided in Supplementary Table 5. Three alignment files including all DNA barcode sequences are available on Figshare[40]. More detailed voucher information, including specimen records from TAIF and links to voucher images, is provided in the GBIF occurrence dataset[41], in which DNA barcode sequences are also gathered. The raw reads resulting from Illumina MiSeq and PacBio libraries had been deposited in NCBI Sequence Read Archive (SRA)[42].

Technical Validation

The ML trees generated from the two DNA barcodes are available on Figshare[40] and Fig. 2. These phylogenies provided strong resolution and branch supports identifying the systematic placement of 1,496 fern samples (Fig. 2), and aligned well with modern classification of ferns[1,43–45]. From the family-level backbone, the combined *rbcL* + *trnL-F* tree shows over 90% of nodes with UFBS values above 90. Individually, the *rbcL* and *trnL-F* trees resolved 84% and 67% of nodes with similarly high supports. At the intergeneric level, the *rbcL* + *trnL-F* phylogeny resolved 89% of nodes with UFBS values above 90, while single-barcode *rbcL* and *trnL-F* trees resolved 83% and 81% with such high support. At the infrageneric level, the *rbcL* + *trnL-F* tree supported the monophyly of 75% of species with multiple collections, after we excluded hybrids and species-unidentified samples. On this scale, the *rbcL* and *trnL-F* trees respectively supported monophyly in 72% and 73% of species.

Code availability

The customized shell script for demultiplexing and removal of primer sequences within NGS reads using cutadapt (v.3.5), and the R (v.4.2.0) script for the generation of dada2's ASV are available at https://github.com/lykuofern/1.5KP_datapaper. The pipeline for technical validation was designed using the 'targets' R package[46] and is available from https://github.com/joelnitta/bifa_barcodes. A Docker image to run the code is available from https://hub.docker.com/r/joelnitta/bifa_barcodes.

Received: 29 April 2024; Accepted: 20 November 2024;
Published online: 02 December 2024

References

1. PPG I. A community-derived classification for extant lycophytes and ferns. *J. Syst. Evol.* **54**, 563–603 (2016).
2. Kreft, H., Jetz, W., Mutke, J. & Barthlott, W. Contrasting environmental and regional effects on global pteridophyte and seed plant diversity. *Ecography* **33**, 408–419 (2010).
3. Moran, R. C. Diversity, biogeography, and floristics. in *Biology and Evolution of Ferns and Lycophytes* (eds. Ranker, T. A. & Haufler, C. H.) 367–394 (Cambridge University Press, 2008).
4. Suissa, J. S., Sundue, M. A. & Testo, W. L. Mountains, climate and niche heterogeneity explain global patterns of fern diversity. *J. Biogeogr.* **48**, 1296–1308 (2021).
5. Kress, W. J., García-Robledo, C., Uriarte, M. & Erickson, D. L. DNA barcodes for ecology, evolution, and conservation. *Trends Ecol. Evol.* **30**, 25–35 (2015).
6. de Groot, G. A. *et al.* Use of *rbcL* and *trnL-F* as a two-locus DNA barcode for identification of NW-European ferns: an ecological perspective. *PLoS One* **6**, e16371 (2011).
7. Wu, Y.-H., Ke, Y.-T., Chan, Y.-Y., Wang, G.-J. & Kuo, L.-Y. Integrating tissue-direct PCR into genetic identification: an upgraded molecular ecology way to survey fern field gametophytes. *Appl. Plant Sci.* **10**, e11462 (2022).
8. Li, F.-W. *et al. rbcL* and *matK* earn two thumbs up as the core DNA barcode for ferns. *PLoS One* **6**, e26597 (2011).
9. Kuo, L.-Y. *et al.* Organelle genome inheritance in *Deparia* ferns (Athyriaceae, Aspleniineae, Polypodiales). *Front. Plant Sci.* **9**, Article 486 (2018).
10. Ebihara, A. & Kuo, L.-Y. East and Southeast Asian pteridophyte flora and DNA barcoding. in *The Biodiversity Observation Network in the Asia-Pacific Region: Toward Further Development of Monitoring, Ecological Research Monographs*. (eds. Nakano, S., Yahara, T. & Nakashizuka, T.) 321–327, https://doi.org/10.1007/978-4-431-54032-8 (Springer Japan, 2012).
11. Ebihara, A., Nitta, J. H. & Ito, M. Molecular species identification with rich floristic sampling: DNA barcoding the pteridophyte flora of Japan. *PLoS One* **5**, e15136 (2010).
12. Nitta, J. H., Ebihara, A. & Smith, A. R. A taxonomic and molecular survey of the pteridophytes of the Nectandra Cloud Forest Reserve, Costa Rica. *PLoS One* **15**, e0241231 (2020).
13. Nitta, J. H., Meyer, J.-Y., Taputuarai, R. & Davis, C. C. Life cycle matters: DNA barcoding reveals contrasting community structure between fern sporophytes and gametophytes. *Ecol. Monogr.* **87**, 278–296 (2017).

14. Trujillo-Argueta, S., del Castillo, R. F., Tejero-Diez, D., Matias-Cervantes, C. A. & Velasco-Murguía, A. DNA barcoding ferns in an unexplored tropical montane cloud forest area of southeast Oaxaca, Mexico. *Sci. Rep.* **11**, 22837 (2021).
15. Nitta, J. H., Schuettpelz, E., Ramírez-Barahona, S. & Iwasaki, W. An open and continuously updated fern tree of life. *Front. Plant Sci.* **13**, 1–17 (2022).
16. Chen, C.-W. *et al*. *TrnL-F* is a powerful marker for DNA identification of field vittarioid gametophytes (Pteridaceae). *Ann. Bot.* **111**, 663–673 (2013).
17. CBOL Plan tWorking Group. A DNA barcode for land plants. *Proc. Natl. Acad. Sci. USA* **106**, 12794–12797 (2009).
18. Li, F.-W., Kuo, L.-Y., Pryer, K. M. & Rothfels, C. J. Genes translocated into the plastid inverted repeat show decelerated substitution rates and elevated GC content. *Genome Biol. Evol.* **8**, 2452–2458 (2016).
19. Kuo, L.-Y., Li, F.-W., Chiou, W.-L. & Wang, C.-N. First insights into fern *matK* phylogeny. *Mol. Phylogenet. Evol.* **59**, 556–566 (2011).
20. Nitta, J. H. & Chambers, S. M. Identifying cryptic fern gametophytes using DNA barcoding: a review. *Appl. Plant Sci.* **10**, e11465 (2022).
21. Ebihara, A. *et al*. A survey of the fern gametophyte flora of Japan: frequent independent occurrences of noncordiform gametophytes. *Am. J. Bot.* **100**, 735–743 (2013).
22. Quinlan, A. *et al*. Providing the missing links in fern life history: insights from a phenological survey of the gametophyte stage. *Appl. Plant Sci.* **10**, e11473 (2022).
23. Farrar, D. R. Independent fern gametophytes in the wild. *Proc. R. Soc. Edinburgh. Sect. B. Biol. Sci.* **86**, 361–369 (1985).
24. Kuo, L.-Y. *et al*. Not only in the temperate zone: independent gametophytes of two vittarioid ferns (Pteridaceae, Polypodiales) in East Asian subtropics. *J. Plant Res.* **130**, 255–262 (2017).
25. Pinson, J. B., Chambers, S. M., Nitta, J. H., Kuo, L.-Y. & Sessa, E. B. The separation of generations: biology and biogeography of long-lived sporophyteless fern gametophytes. *Int. J. Plant Sci.* **178**, 1–18 (2017).
26. Wizenberg, S. B. *et al*. Environmental metagenetics unveil novel plant-pollinator interactions. *Ecol. Evol.* **13**, 1–8 (2023).
27. Chang, J. S., Lee, S. Y. & Kim, K. W. Arsenic in an As-contaminated abandoned mine was mobilized from fern-rhizobium to frond-bacteria via the ars gene. *Biotechnol. Bioprocess Eng.* **15**, 862–873 (2010).
28. Tschudy, R. H., Pillmore, C. L., Orth, C. J., Gilmore, J. S. & Knight, J. D. Disruption of the terrestrial plant ecosystem at the Cretaceous-Tertiary boundary, western interior. *Science* **225**, 1030–1032 (1984).
29. Azevedo-Schmidt, L. *et al*. Ferns as facilitators of community recovery following biotic upheaval. *Bioscience* biae022, https://doi.org/10.1093/biosci/biae022 (2024).
30. Deiner, K. *et al*. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.* **26**, 5872–5895 (2017).
31. Kuo, L.-Y. Polyploidy and biogeography in genus *Deparia* and phylogeography in *Deparia lancea*. (National Taiwan University, Ph.D thesis, 2015).
32. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
33. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10 (2011).
34. Callahan, B. J. *et al*. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
35. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
36. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
37. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
38. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
39. Nguyen, L. T., Schmidt, H. A., VonHaeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
40. Kuo, L.-Y. & Nitta, J. H. Alignments and ML phylogenies of DNA barcodes of 1496 Asian fern samples. *Figshare* https://doi.org/10.6084/m9.figshare.27649653 (2024).
41. Chen, C.-W. & Kuo, L.-Y. Improving knowledge of Asian pteridophytes through DNA sampling of specimens in regional collections. Version 1.5. Taiwan Forestry Research Institute. *GBIF.org* https://doi.org/10.15468/nhsnn2 (2024).
42. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP499440 (2024).
43. Zhou, X.-M. *et al*. A global phylogeny of grammitid ferns (Polypodiaceae) and its systematic implications. *Taxon* **72**, 974–1018 (2023).
44. Fawcett, S. *et al*. A global phylogenomic study of the Thelypteridaceae. *Syst. Bot.* **46**, 891–915 (2021).
45. Chen, C.-C., Hyvönen, J. & Schneider, H. Exploring phylogeny of the microsoroid ferns (Polypodiaceae) based on six plastid DNA markers. *Mol. Phylogenet. Evol.* **143**, 106665 (2020).
46. Landau, W. The targets R package: a dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *J. Open Source Softw.* **6**, 2959 (2021).
47. Vasco, A., Moran, R. C. & Ambrose, B. A. The evolution, morphology, and development of fern leaves. *Front. Plant Sci.* **4**, Article 345 (2013).
48. Dong, S.-Y. *et al*. New species of the fern genus *Lindsaea* (Lindsaeaceae) from New Guinea with notes on the phylogeny of *L*. sect. *Synaphlebium*. *PLoS One* **11**, e0163686 (2016).

Acknowledgements

We thank Fay-Wei Li, Hsin-Chieh Hung, Min-Han Pan, Nguyễn Khánh Trình Trầm, Pei-Hsuan Lee, Pi-Fong Lu, Yea-Chen Liu, You-Wun Hwang, Sheng-Sian Dai, Wei-Ting Liou, Wen-Liang Chiou, and Yao-Moan Huang for help in specimen collections; Ya-Ting Ke, Yu-Chia Wang, and Yu-Hong Li for their help with molecular work; Alexandria Quinlan for comments for the research grant proposal; Melissa Jean-Yi Liu for the assistance of the GBIF dataset; two anonymous reviewers for their comments for the manuscript; Taiwan Pteridophyte Research Group (TPG) for maintaining DNA samples; Alan R. Smith for providing initial identification for Thelypteridaceae specimens; Genomics Corp. (New Taipei City, Taiwan), Health GeneTech Corp. (New Taipei City, Taiwan), and Sequencing and Genomic Technologies Core Facility of the Duke University Center for Genomic and Computational Biology (Durham, North Carolina, USA) for sequencing. This study was supported by Ministry of the Environment (Government of Japan) under Biodiversity Information Fund for Asia project (BIFA6_010; https://www.gbif.org/project/BIFA6_010/), MOST project (109-2621-B-007-001-MY3, 111-2628-B-007-006-MY3) in Taiwan, and the U.S. National Science Foundation (award DEB-1701942 to Kathleen Pryer at Duke University and T.-T.K.).

Author contributions

L.-Y.K. designed this study and managed the progress of the project; L.-Y.K., C.-Wei C., Z.-X.C., T.-C.H., Y.-H.C., Y.-S.C., H.T.L., T.-T.K., A.M.A.M., F.P.C., V.B.A., and Y.K.T. conducted the field works and collected specimens and the DNA samples; L.-Y.K., C.-Wei C., Z.-X.C., T.-C.H., Y.-H.C., Y.-S.C., S.F., and M.S. identified the specimens; L.-Y.K., Y.-H.H., P.-J.X., and T.-T.K. performed the experiments; L.-Y.K., S.-K.T., C.-Wei C., J.H.N. complied the datasets; J.H.N. and L.-Y.K. analysed the data; L.-Y.K. prepared the draft of manuscript with the significant inputs from M.S., T.-T.K., S.F. and J.H.N.

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/s41597-024-04161-8.

Correspondence and requests for materials should be addressed to L.-Y.K.

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.