A Comprehensive Phylogenomic Platform for Exploring the Angiosperm Tree of Life

William J. Baker[1,*], Paul Bailey[1], Vanessa Barber[1], Abigail Barker[1], Sidonie Bellot[1], David Bishop[1], Laura R. Botigué[1,2], Grace Brewer[1], Tom Carruthers[1], James J. Clarkson[1], Jeffrey Cook[1], Robyn S. Cowan[1], Steven Dodsworth[1,3], Niroshini Epitawalage[1], Elaine Françoso[1], Berta Gallego[1], Matthew G. Johnson[4], Jan T. Kim[1,5], Kevin Leempoel[1], Olivier Maurin[1], Catherine McGinnie[1], Lisa Pokorny[1,6], Shyamali Roy[1], Malcolm Stone[1], Eduardo Toledo[1], Norman J. Wickett[7], Alexandre R. Zuntini[1], Wolf L. Eiserhardt[1,8,†], Paul J. Kersey[1,†], Ilia J. Leitch[1,†], Félix Forest[1,†]

[1]Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AE, United Kingdom

[2]Current address: Centre for Research in Agricultural Genomics, Campus UAB, Edifici CRAG, Bellaterra Cerdanyola del Vallès, 08193 Barcelona, Spain

[3]School of Life Sciences, University of Bedfordshire, University Square, Luton LU1 3JU, United Kingdom

[4]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA

[5]Current address: Department of Computer Science, School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, United Kingdom

[6]Current address: Centre for Plant Biotechnology and Genomics (CBGP) UPM-INIA, 28223 Pozuelo de Alarcón (Madrid), Spain

[7]Plant Science and Conservation, Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022, USA

[8]Department of Biology, Aarhus University, 8000 Aarhus C, Denmark

†Joint senior authors

*Corresponding author: Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AE, United Kingdom, w.baker@kew.org

Abstract.—The tree of life is the fundamental biological roadmap for navigating the evolution and properties of life on Earth, and yet remains largely unknown. Even angiosperms (flowering plants) are fraught with data gaps, despite their critical role in sustaining terrestrial life. Today, high-throughput sequencing promises to significantly deepen our understanding of evolutionary relationships. Here, we describe a comprehensive phylogenomic platform for exploring the angiosperm tree of life, comprising a set of open tools and data based on the 353 nuclear genes targeted by the universal Angiosperms353 sequence capture probes. This paper (i) documents our methods, (ii) describes our first data release and (iii) presents a novel open data portal, the Kew Tree of Life Explorer (https://treeoflife.kew.org). We aim to generate novel target sequence capture data for all genera of flowering plants, exploiting natural history collections such as herbarium specimens, and augment it with mined public data. Our first data release, described here, is the most extensive nuclear phylogenomic dataset for angiosperms to date, comprising 3,099 samples validated by DNA barcode and phylogenetic tests, representing all 64 orders, 404 families (96%) and 2,333 genera (17%). Using the multi-species coalescent, we inferred a "first pass" angiosperm tree of life from the data, which totalled 824,878 sequences, 489,086,049 base pairs, and 532,260 alignment columns. The tree is strongly supported and highly congruent with existing taxonomy, while challenging numerous hypothesized relationships among orders and placing many genera for the first time. The validated dataset, species tree and all intermediates are openly accessible via the Kew Tree of Life Explorer. This major milestone towards a complete tree of life for all flowering plant species opens doors to a highly integrated future for angiosperm

phylogenomics through the systematic sequencing of standardised nuclear markers. Our approach has the potential to serve as a much-needed bridge between the growing movement to sequence the genomes of all life on Earth and the vast phylogenomic potential of the world's natural history collections.

Keywords: angiosperms, Angiosperms353, genomics, herbariomics, museomics, nuclear phylogenomics, open access, target sequence capture, tree of life.

INTRODUCTION

Discovering the tree of life is among the most fundamental of the grand challenges in science today (Hinchliff et al. 2015). The tree of life is the biological roadmap that allows us to discover, identify and classify life on Earth, to explore its properties, to understand its origins and evolution, and to predict how it will respond to future environmental change. Of all eukaryotic lineages, the angiosperms (flowering plants) are among the most pressing priorities for tree of life research. Angiosperms sustain the terrestrial living world, including humanity, as primary producers, ecosystem engineers and earth system regulators. They hold potential solutions to global challenges, such as climate change, biodiversity loss, human health, food security and renewable energy (Antonelli et al. 2020). In light of this, a phylogenetic framework with which to navigate and interpret the species, trait and functional diversity of angiosperms has never been more necessary. However, despite substantial progress, the evolutionary connections among Earth's ca. 330,000 flowering plant species (WCVP 2020) remain incompletely known.

The angiosperm research community were early and organised adopters of the molecular phylogenetic approach, resulting in numerous benchmark tree of life publications (e.g. Chase et al. 1993; Soltis et al. 2008; Soltis et al. 2011), and a community approach to

phylogenetic classification (APG 1998; APG II 2003; APG III 2009; APG IV 2016). Through this distributed effort, a wealth of DNA sequence data is now available in public repositories, covering ca. 107,000 (31%) of the ca. 350,000 species of vascular plants (RBG Kew 2016; WCVP 2020), most of which are angiosperms (see also Cornwell et al. 2019). However, the lack of sequence data for the remaining 69% obstructs their accurate placement in the tree of life. In addition, lack of complementarity in gene sampling across public DNA sequence data impedes phylogenetic synthesis (Hinchliff and Smith 2014). For example, data from either one or both of *rbcL* and *matK*, the two most popular plastid genes for phylogenetics, are available for only 54% of the ca. 107,000 sequenced vascular plant species (RBG Kew 2016). Comprehensive phylogenetic trees of flowering plants are in high demand (Hinchliff et al. 2015; Eiserhardt et al. 2018), but currently can only be made "complete" using proxies, such as taxonomic classification, to interpolate the unsequenced species (Smith and Brown 2018), which may not accurately reflect relationships. Greater community-wide coordination of both taxon and gene sampling would benefit phylogenetic data integration immensely, creating numerous downstream scientific opportunities.

High-throughput sequencing (HTS) now promises to significantly deepen our understanding of evolutionary relationships among Earth's species, including angiosperms (Li et al. 2019; Yang et al. 2020). For example, the One Thousand Plant Transcriptomes (1KP) initiative has brought an unprecedented scale of data to bear on the plant tree of life (Wickett et al. 2014; Gitzendanner et al. 2018; Leebens-Mack et al. 2019). Nevertheless, with greatly increased data depth come trade-offs in taxon sampling; the pre-eminent HTS studies cited here account for less than 0.01% of angiosperm species. Undeterred by this sampling gap, the Earth Biogenome Project (EBP) has launched a "moonshot for biology" by proposing to sequence and characterise the genomes of all of Earth's eukaryotic species over a 10 year period (Lewin et al. 2018). Projects such as the 10,000 Plant Genomes Project

(Cheng et al. 2018) and the Darwin Tree of Life Project (https://www.darwintreeoflife.org/) aim to contribute to this goal by producing numerous chromosome-level genome assemblies across major lineages and regional biotas. However, taxon sampling remains a significant issue, due to the challenges of obtaining the high molecular weight DNA required by these projects (for long-read HTS) from samples that are both authentically identified and compliant with the spirit and letter of the Nagoya Protocol (Secretariat of the Convention on Biological Diversity 2011). Despite its immense potential, the "whole genome" approach to discovering the tree of life remains a future goal that will not be achieved on a large taxonomic scale in the short term. Methodological compromises are required to accelerate progress.

The world's natural history collections are a goldmine for genomic research (Buerki and Baker 2016), containing tissues of almost all species of life on Earth known to science. However, the condition of these tissues and the DNA therein varies widely, depending on age and preservation techniques, among other factors. In the case of plants, herbarium specimens generally yield degraded DNA, which, though not useful for long-read HTS, is now being intensively exploited for short-read HTS (Bakker et al. 2016; Brewer et al. 2019; Forrest et al. 2019; Alsos et al. 2020). In this context, target sequence capture is growing in popularity as the HTS method most widely applied to herbarium DNA (Dodsworth et al. 2019). This approach (also known as target enrichment, target capture, sequence capture, anchored hybrid enrichment) and its variations (e.g. Hyb-Seq, which combines target sequence capture with genome skimming) use RNA or DNA probes to enrich sequencing libraries for specifically targeted loci (Faircloth et al. 2012; Lemmon et al. 2012; Weitemier et al. 2014). It is proving to be an increasingly cost-effective means of isolating hundreds of loci for phylogenetic analysis from even centuries-old specimens (Brewer et al. 2019), bringing comprehensive

taxon sampling from herbarium collections within the reach of any phylogenomic researcher (Hale et al. 2020).

Numerous target sequence probe sets have been developed for specific angiosperm groups (e.g. Annonaceae [Couvreur et al. 2019], Asteraceae [Mandel et al. 2014], *Dioscorea* [Soto Gomez et al. 2019], *Euphorbia* [Villaverde et al. 2018]). The design of these probe sets is informed by available genomic resources, as well as criteria specific to the group of interest and research questions. As a result, locus overlap between probe sets tends to be minimal. Unlike the Sanger sequencing era, in which researchers converged on tractable genes such as *rbcL* and *matK*, the lack of complementarity between probe sets curtails prospects for data integration across broad taxonomic scales. In addition, development of custom probe sets is expensive, requiring considerable genomic resources and bioinformatic expertise. A publicly available, universal probe set for angiosperms targeting a standard set of loci would resolve these issues (Buddenhagen et al. 2016; Chau et al. 2018). In response to this, we designed the Angiosperms353 probe set (Johnson et al. 2019), drawing on 1KP transcriptome data from ca. 650 angiosperm species (Leebens-Mack et al. 2019). The probe set targets 353 genes from 410 low-copy, protein-coding nuclear orthologs previously selected for phylogenetic analysis across green plants (Leebens-Mack et al. 2019), enriching up to ca. 260 kbp from any flowering plant. Angiosperms353 probes are an open data resource that can be used without the expense of design or access to prior genomic data and have already been successfully applied across different taxonomic scales (e.g. Larridon et al. 2019; Murphy et al. 2020; Pérez-Escobar et al. 2020; Shee et al. 2020), including at the population level (Van Andel et al. 2019; Slimp et al. 2020; Beck et al. 2021).

Here, we describe a large-scale effort to establish a new phylogenomic platform for exploring the angiosperm tree of life, comprising a set of open tools (Angiosperms353 probes, laboratory protocols, analysis pipeline, data portal) and data (sequence data,

assembled genes, alignments, gene trees, species tree). This platform, which directly addresses the challenges outlined above, is an outcome of the Plant and Fungal Trees of Life project (PAFTOL; www.paftol.org) at the Royal Botanic Gardens, Kew (RBG Kew 2015). As a step towards the ultimate goal of a complete species-level tree, we aim to gather DNA sequence data for the Angiosperms353 genes from one species of all 13,862 angiosperm genera (WCVP 2020). This unprecedented dataset of standard loci draws extensively on herbarium collections for comprehensive sampling, especially of genera that have not been sequenced before (Brewer et al. 2019). Extensive new data have been generated, analysed and released into the public domain, along with corresponding phylogenetic inferences. By providing our data in open and accessible ways, including an interactive tree of life, we aim to foster a transparent and collaborative environment for future data re-use and synthesis. This paper serves as the baseline reference for our platform, (i) documenting our methods, (ii) describing our first data release, comprising 17% of angiosperm genera, including initial insights on phylogenetic performance, and (iii) presenting a novel data portal, the Kew Tree of Life Explorer, through which our data and corresponding tree of life can be interrogated and downloaded. We conclude with reflections on the prospects for our approach, future development requirements and the role of open data for enhancing cross-community collaboration towards a complete tree of life.

MATERIALS AND METHODS

This section describes the workflow (Fig. 1) used by the PAFTOL project to generate our first full data release (i.e. Data Release 1.0), which is publicly accessible through our open data portal, the Kew Tree of Life Explorer (https://treeoflife.kew.org), described below. The workflow consists of three main stages: (i) sample processing, encompassing sample

selection and laboratory protocols for target sequence capture data generation (Fig. 2), (ii) data analysis, including target gene assembly, data mining, data validation and phylogenetic inference (Figs. 3, 4), and (iii) data publication via the data portal (Fig. 5). The data accessible via the portal comprise raw data (unprocessed sequence reads) and results from "first pass" analyses (gene assemblies, alignments, gene trees, species tree). Though not exhaustive, these first explorations of the data apply methods that are both rigorous and tractable at our scale of operation.

Details of the first data release are also given in the data release notes in the portal via our secure FTP (http://sftp.kew.org/pub/treeoflife/) and are also archived at the Royal Botanic Gardens, Kew (RBGK) Research Repository (https://doi.org/10.34885/paftol). A new release note will be published in the same locations with each future data release and will detail any changes in methods used relative to the first release described here.

Sampling

We aimed to generate novel data from across the angiosperms, using a stratified sampling approach of one species per genus. Our sampling was standardised to the complete list of angiosperms within the World Checklist of Vascular Plants (WCVP 2020), which currently recognises 13,862 accepted genera in 418 families, aligned to the 64 orders of the APG IV classification (APG IV 2016). We prioritised genera that were not represented by published transcriptomic or genomic data in public sequence repositories (e.g. GenBank), and avoided genera that had already been sampled in large genomic initiatives such as the 1KP project (Leebens-Mack et al. 2019). The selection of species within genera was made pragmatically, although we prioritised the species of the generic type where possible.

Plant material was obtained from a variety of sources (Fig. 2), primarily from the collections of RBGK (herbarium, DNA bank, silica gel-dried tissue collection, living collection and the Millennium Seed Bank, https://www.kew.org/science/collections-and-

resources/collections). Additional material (tissue samples, extracted DNA) was generously provided by our collaborative networks (see Acknowledgements). To be selected, the material must have been (i) legally sourced and made available for use in phylogenomic studies, (ii) identified to species level, preferably by an expert of the group, and (iii) ideally collected in the wild. As far as was practically achievable, we ensured that the identity of each sample was substantiated by a voucher specimen deposited in a publicly accessible herbarium.

All metadata were captured using a relational database that allowed us to track processing of samples from the selection of material, through the library preparation pipeline to the completion of sequencing. Data were recorded in four main tables (Specimen, Sample, Library, Sequencing). The database architecture allowed us to record multiple sequence datasets (fastq files) from one or several libraries, and one or several DNA extracts from a single specimen. Relevant voucher specimen information was also captured in the database (e.g. collector(s), collector number, herbarium acronym (following Index Herbariorum http://sweetgum.nybg.org/science/ih/), country of origin, date of collection, specimen barcodes). Voucher data are available via our data portal (see below). Images of specimens sampled from the RBGK Herbarium are in the process of being captured in RBGK's online herbarium catalogue (http://apps.kew.org/herbcat/) and, where available, are linked to the appropriate records in the Kew Tree of Life Explorer.

DNA extraction

DNA was extracted from 40 mg of herbarium material, 20 mg of silica gel-dried material (Chase and Hills 1991), or 100 mg of fresh material using a modified CTAB extraction method (Doyle and Doyle 1987; Fig. 2). Plant tissue was pulverized using a Mixer Mill MM400 (Retsch GmbH, Germany). DNA extractions were purified by a magnetic bead

clean-up using Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN, USA), according to the manufacturer's protocols. Samples obtained from the RBGK DNA bank (http://dnabank.science.kew.org/homepage.html) had been extracted using a modified CTAB method (Doyle and Doyle 1987) followed by caesium chloride/ethidium bromide density gradient cleaning and dialysis. DNA samples provided by external collaborators had been extracted using a wide variety of extraction methods from living, silica gel-dried and herbarium material.

All DNA samples were quality checked for concentration and degree of fragmentation. DNA concentration was measured using a Quantus (Promega, Madison, WI, USA) or Qubit (Thermo Fisher Scientific, Inchinnan, UK) fluorometer. DNA fragment size range was routinely assessed on a 1% agarose gel using ethidium bromide and visualized with a UVP Gel Studio (AnalytikJena, Jena, Germany). For samples with a low DNA concentration (i.e. not visible on a gel), fragment sizes were assessed on a 4200 TapeStation using Genomic DNA ScreenTape (Agilent Technologies, Cheadle, UK).

Library preparation

Genomic DNA samples were diluted to 4 ng/µl with 10 mM Tris (pH 8.0). Those with an average fragment size greater than 350 bp were sonicated to an average fragment size ca. 400 bp, using a Covaris M220 Focused-ultrasonicator (Covaris, Woburn, MA, USA) by adding 50 µl of diluted genomic DNA to a 130 µl Covaris microAFA tube. The sonication time was adjusted for each sample based on its average DNA fragment size (15 to 100 secs, following the manufacturer's protocols). Additional parameters used were peak incident power to 50W, duty factor to 10% and 200 cycles per burst.

Libraries were prepared using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, Ipswich, MA, USA; Fig. 2). Size selection was not employed for samples with highly degraded DNA. In the early stages of the project, libraries were prepared

following the manufacturer's protocols exactly, but the majority were prepared using half of the recommended volumes throughout to increase cost efficiency. All DNA fragments were indexed using NEBNext Multiplex Oligos for Illumina (Dual Index Primer sets 1 and 2, New England Biolabs, Ipswich, MA, USA).

The distribution of fragment sizes in each library was assessed with a 4200 TapeStation using standard D1000 tapes. Library concentration was measured using a Quantus fluorometer. If the library concentration was less than 10 nM, up to eight additional PCR cycles were performed, following the NEBNext Ultra II Library Prep Kit protocol with IS5_reamp.P5 and IS6_reamp.P7 primers (Meyer and Kircher 2010). Library quality assessment was then repeated.

Pooling and hybridisation

Prior to hybridisation (Fig. 2), all libraries were normalised to 10 nM, using 10 mM Tris (pH 8.0) and then combined into pools of 20 to 24 libraries, each containing 10 µl (0.1 pmol) of each normalized library (i.e. a total of ca. 600-700 ng DNA in each pool, assuming an average fragment size of ca. 450 bp). To ensure even sequencing across all samples in a pool, species for pooling were selected to minimize the range of DNA fragment sizes and ensure a narrow taxonomic breadth. The latter criterion was needed because samples that are more closely related to the taxa used to construct the probe set tend to preferentially hybridise. This can lead to an over-representation of their sequences in the DNA data if appropriate care is not taken when selecting species for the sequencing pool. In rare cases, such as smaller pools (ca. 10 libraries) of short fragment (i.e. <300 bp) libraries, it was necessary to recalculate the standard volume of normalized libraries to be added to ensure that the final pool contained ca. 500 ng of DNA.

The pooled libraries were dried in a SpinVac (Eppendorf, Dusseldorf, Germany), resuspended in 8 µl of 10 mM Tris (pH 8.0) and enriched by hybridising with the

Angiosperms353 probe kit (Johnson et al. 2019; Arbor Biosciences myBaits Target Sequence Capture Kit, 'Angiosperms 353 v1', Catalogue #308196) following the manufacturer's protocol, version 4.0. Hybridisation was typically performed at 65ºC for 24 h, with reactions topped with 30 µl of red Chill-out Liquid Wax (Bio-Rad, Hercules, CA, USA) to prevent evaporation. However, for short libraries (i.e. <350 bp) the temperature was reduced to 60ºC, following the recommendations of Arbor Biosciences.

The target-enriched pools were amplified using the KAPA HiFi 2X HotStart ReadyMix PCR Kit (Roche, Basel, Switzerland) or NEBNext Q5 HotStart HiFi PCR Master Mix (New England BioLabs, Ipswich, MA, USA) for eight to 14 cycles. Amplified pools were then purified using Agencourt AMPure XP Beads (at 0.9X the sample volume) and eluted in 15 µl of 10 mM Tris (pH 8.0).

Products were quantified with a Quantus fluorometer and re-amplified if the concentration was below 6 nM, with three to six PCR cycles (see above). Final products were assessed using the TapeStation to determine the distribution of fragment sizes. The target-enriched pools were normalized to 6 nM (using 10 nM Tris, pH 8.0) and multiplexed for sequencing, with the number of target-enriched pools combined in each sequencing pool varying from two to 20 (comprising a total of 48-384 samples) depending on the sequencing platform and service provider requirements.

DNA sequencing

Initially, DNA sequencing was performed on an Illumina MiSeq at RBGK with version 3 chemistry (Illumina, San Diego, CA, USA) and ran for 600 cycles to generate 2 × 300 bp paired-end reads. Subsequently, DNA sequencing was outsourced (Macrogen, Seoul, South Korea, or Genewiz, Takeley, UK) and performed on an Illumina HiSeq producing 2 × 150 bp paired-end reads. Raw reads were deposited in the European Nucleotide Archive

297 under an umbrella project (accession number PRJEB35285) and can be accessed from the
298 individual sample records in the Kew Tree of Life Explorer.
299
300 **Sequence assembly**
301 Coding sequences were recovered from target-enriched sequence data using our
302 pipeline recoverSeqs (accessible from our GitHub repository
303 https://github.com/RBGKew/KewTreeOfLife, pypaftol 'paftools' submodule) to retrieve
304 sequences orthologous to the Angiosperms353 target gene set (Johnson et al. 2019;
305 https://github.com/mossmatters/Angiosperms353). This target set contained multiple
306 reference sequences per gene, thereby covering a large phylogenetic breadth to facilitate read
307 recovery across angiosperms.
308 The process comprised four main stages (Fig. 3), applied to each sample**:** (i) sequence
309 reads were trimmed using Trimmomatic (Bolger et al. 2014) with the following parameters:
310 ILLUMINACLIP: <AdapterFastaFile>: 2:30:10:2:true, LEADING: 10, TRAILING: 10,
311 SLIDINGWINDOW: 4:20, MINLEN: 40, with the adaptor fasta file formatted for
312 palindrome trimming, (ii) trimmed read pairs were mapped to the Angiosperms353 target
313 genes with TBLASTN. A representative reference sequence for each gene was then selected
314 by identifying the sequence with the largest number of mapped reads. (iii) This representative
315 gene was used as the reference for assembling the gene-specific reads using an overlap-based
316 assembly algorithm (--assembler overlapSerial option) as follows. First, the reads were
317 aligned to and ordered along the reference sequence based on a minimum alignment size of
318 50 bases (--windowSizeReference option) with a minimum sequence identity of 70% (--
319 relIdentityThresholdReference option). Consecutive reads ordered along the reference
320 sequence were aligned in a pair-wise manner to find read overlaps. If an overlap of at least 30
321 bases (--windowSizeReadOverlap option) and 90% sequence identity (--

relIdentityThresholdReadOverlap option) was found, the aligned reads were used to construct a consensus contig with ambiguous bases represented by 'N'. This last parameter resulted in one or more sets of aligned reads with ≥90% sequence identity, each set being merged into a single contig. In the final stage, the exonerate protein2genome program was used to identify the exon-intron structure within each contig. One or more contigs were chosen that best represented the structure of the exon(s) in the reference gene chosen in step (ii). If the exons existed in multiple contigs, those contigs were joined together to form the recovered gene coding sequence.

Target gene recovery success was assessed for each sample by calculating the number of genes recovered and the sum of the recovered gene lengths. Samples were removed from downstream analyses if the sum of the recovered gene lengths fell below 20% of the median value across all samples.

Public data mining

In addition to newly generated target sequence capture data, the Angiosperms353 genes were mined from publicly available genomic data (Fig. 3). For the first release, we mined data from the 1KP Initiative (Carpenter et al. 2019; Leebens-Mack et al. 2019) and published genomes with gene annotations (https://plants.ensembl.org/). The genes were retrieved from assembled transcript sequences (1KP) or coding sequences (CDS; genomes) using paftools retrievetargets from our pipeline, which relies on BLASTN to identify and extract the genomic or transcriptomic sequences corresponding to the 353 genes. Because initial recovery of genes from 1KP transcripts was unsatisfactory, we expanded the Angiosperms353 target set (dataset available from our GitHub) to improve matching and retrieval of genes. As with the novel target sequence capture assemblies, data were removed

from downstream analyses if the sum of the gene lengths fell below 20% of the median value across all samples.

Family identification validation

To verify the family identification of our processed samples, we implemented two validation steps, which were run in parallel (Fig. 4). The two steps consisted of (i) DNA barcode validation, which utilised nuclear ribosomal and plastid barcodes for DNA-based identification, and (ii) phylogenetic validation, which checked the placement of each sample in a preliminary tree relative to its expected position based on its initial family assignment. Identification checks below the family level were not conducted due to the incompleteness of adequate reference resources for DNA barcode validation and sparseness of sampling for phylogenetic validation at the genus or species level.

For barcode validation of target sequence capture data (Fig. 4), plastomes and ribosomal DNA were recovered from raw reads using GetOrganelle (Jin et al. 2020) and subsequently queried against databases of reference plant barcodes using BLASTN (Camacho et al. 2009). For 1KP samples, transcriptome assemblies were directly used as queries in BLASTN. Note that we considered the family identity of annotated genomes to be correct and hence a barcode validation was unnecessary. Six individual barcode reference databases were built from the NCBI nucleotide and BOLD databases (https://www.ncbi.nlm.nih.gov/nuccore; https://www.boldsystems.org/, accessed on 29/10/2020), one for the whole plastome, and the remaining five for specific loci (nuclear ribosomal 18S, as well as plastid *rbcL*, *matK*, *trnL*, and *trnH-psbA*). As for samples, the taxonomy of reference sequences was standardized to WCVP (WCVP 2020). BLAST results were further filtered with a minimum identity >95% and a minimum coverage of reference

locus ≥90% (except for whole plastomes, for which only a filtering based on minimum length was applied).

Tests could only be completed if a sample's given family was present in the barcode databases and if at least one BLAST match remained after filtering. Thus, zero to six barcode tests were conducted per sample. A sample passed an individual test if the first ranked BLAST match (ranked by percentage of identity) confirmed its original family identification and failed otherwise. The final result of the barcode validation following the six individual barcode tests were determined as follows: (i) Confirmed, if one or more barcode tests matched the family identification of a sample; (ii) Rejected, if more than half of the barcode tests gave the same incorrect family identification (requires at least two barcode tests); (iii) Inconclusive (otherwise). Further details of the barcode validation methods can be found in Supplementary Material available on Dryad. The scripts and lists of NCBI and BOLD accessions used in barcode databases are available on our GitHub repository.

To conduct phylogenetic validation (Fig. 4), a preliminary phylogenetic tree was built using the complete, unvalidated dataset, following the phylogenetic methods described below. We then assessed which nodes best represented each order and family in the tree. For every node in the tree, two metrics were calculated for all families and orders: (i) the proportion of samples belonging to a given order/family that are descendants of the node, and (ii) the proportion of samples descending from the node that belong to the order/family. The two metrics were then multiplied to produce an overall taxon concordance score. For each family and order, the highest scoring node was subsequently considered to best represent the taxon in the tree (allowing the identification of outlying samples). A node with a score of 1 for a given order/family is the crown node (most recent common ancestral node) of that taxon, which is monophyletic in the tree. See Supplementary Figure S1 for an illustration.

The family identification of each sample was determined as (i) Confirmed: if identified as belonging to a family whose best scoring node had a taxon concordance score >0.5 and found as a descendant of this node in the tree, (ii) Rejected: if identified as belonging to a family whose best scoring node had a taxon concordance score >0.5 but not found as a descendant of this node, or (iii) Inconclusive: if identified as belonging to a family whose best scoring node had a taxon concordance score ≤0.5. Note that for families represented in the tree by a single sample, the validation was performed with respect to their orders. If the order was represented by a single sample, the validation result was coded as inconclusive.

The outputs of the phylogenetic and DNA barcode validation were combined to identify samples for automatic inclusion and exclusion from the final dataset, and samples for which a decision on inclusion/exclusion was subject to expert review (Fig. 4). Exclusions after expert review were made based on implausible tree placement (e.g. wrong higher clade) or sample misidentification (e.g. match to another family in the barcode validation).

All assembled Angiosperms353 gene data from all samples validated for inclusion form the basis of Data Release 1.0. These were made publicly available via the Kew Tree of Life Explorer.

Phylogeny estimation

We inferred a phylogenetic tree from all validated data (Data Release 1.0) for presentation in an interactive format in the Kew Tree of Life Explorer. This species tree was estimated from gene trees using the multi-species coalescent summary method implemented in ASTRAL-III (Zhang et al. 2018). In addition to the angiosperm samples, ten samples representing seven gymnosperm families from the 1KP initiative were mined for Angiosperms353 orthologs and included in all analyses as outgroup taxa. Our phylogenomic pipeline, available from our GitHub repository, is summarised below.

For each gene, DNA sequences were aligned with UPP 4.3.12 (Nguyen et al. 2015). At the start of the alignment process a set of 1,000 sequences were selected for an initial backbone tree. Option -M was set to '-1' so that sequences could be selected within 25% of the median full-length sequence. Filtering and trimming of the alignment were performed with AMAS (Borowiec 2016) as follows. Sequences with insufficient coverage (<60%) across well occupied columns of each gene alignment were removed. Well occupied columns were defined as those with more than 70% of positions occupied. Then, alignment columns with <0.3% occupancy were removed to avoid a large number of columns with very rare or unique insertions from being included in the tree reconstruction. Finally, sequences with a total length of less than 80 bases were removed, and genes with <30 overlapping bases (at the 70% threshold mentioned above) were excluded.

Gene trees were estimated with IQ-TREE 2.0.5 (Minh et al. 2020) inferring branch support using the ultrafast bootstrap method (option -B; Hoang et al. 2017) with the maximum number of iterations set to 1,000 (option -nm) and using a single model of evolution (option -m GTR+F+R). The use of a single model without testing many models of evolution was a pragmatic choice, following Abadi et al. (2019). TreeShrink 1.3.4 (Mai and Mirarab 2018) was used to remove abnormally long branches from gene trees using default settings, except option -b, which was set to 20. The alignment and gene tree estimation steps were then repeated on the samples retained by TreeShrink. Before reconstructing the species tree using ASTRAL-III, nodes in the gene trees with bootstrap support values less than 30% were collapsed using nw_ed from Newick Utilities 1.6.0 (Junier and Zdobnov 2010). This value was deduced from interpreting Figure 1 in Hoang et al. (2017), adjusting the standard bootstrap threshold of 10% (recommended for ASTRAL-III), to 30 % for the ultrafast bootstrap.

All gene alignments, gene trees and the ASTRAL-III species tree are available for download from secure FTP and the Kew Tree of Life Explorer. In addition, the species tree is available to browse through an interactive tree viewer implemented within the Kew Tree of Life Explorer (see also Supplementary Fig. S2).

Data portal implementation

To disseminate results, a data portal (the Kew Tree of Life Explorer; https://treeoflife.kew.org) was designed and implemented (Fig. 5) with a layered architecture that comprised: (i) a MariaDB running on a Galera multi-master cluster as a database management system; (ii) an API written in Python using the Flask framework and the SQLAlchemy library; (iii) a front-end written using the Vue.js framework and Nuxt.js for the tabular data (used to provide access to gene and specimen data) and content pages; (iv) a tree visualisation module developed from the open source application PhyD3 (Kreft et al. 2017) using D3.js (Bostock 2012) for data visualisation; and (v) deployment on a Linux (CentOS 7) server using Nginx as web server and load balancer.

The data, with appropriate metadata and documentation, are available for public download over secure FTP (http://sftp.kew.org/pub/treeoflife/) and the Kew Tree of Life Explorer under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. When superseded by new releases, archived earlier releases will remain accessible via secure FTP.

RESULTS

Initial dataset

The initial dataset prior to processing and analysis comprised data from 3,272 angiosperm samples, representing 413 families of angiosperms (99%) and 2,428 genera (18%; Table 1). We generated novel target sequence capture data for 2,522 of these samples,

which included 104 angiosperm genera that have never been sequenced before. Data for the remainder were mined from public sources (689 1KP transcriptomes, 61 annotated genomes). The majority of target sequence capture data were generated from the RBGK collections as follows: DNA Bank (43%), herbarium (28%), silica gel-dried tissue collection (8%), living collection (2%), and Millennium Seed Bank (0.3%). The remaining 19% of samples included in this study were provided by various collaborators of the PAFTOL project, either as DNA samples or as dried tissue (see Acknowledgements).

Sequence recovery from all 2,522 target sequence capture samples (prior to any quality controls) is visualised in Figure 6. Eighty-four target sequence capture samples and eleven 1KP transcriptomes were removed from downstream analyses because the sum of gene lengths did not meet the quality threshold of 20% of the median value across all samples.

Family identification validation

The remaining 3,177 samples (Table 1) were processed through our sample family identification validation pipeline (Fig. 4, Table 2, Supplementary Table S1). Of these, 3,064 (97%) were automatically cleared for inclusion and 67 were automatically excluded (Table 2). The remaining 46 samples were held for expert review, after which 35 were cleared for inclusion and 11 were excluded due to implausible tree placements. The majority of excluded samples (64 out of 78) were from the novel target sequence capture data, although 14 were 1KP transcriptomes, highlighting the risk of sample misidentification in even the most highly curated datasets. Further details regarding the results obtained during the family identification validation by DNA barcoding can be found in Supplementary Material available on Dryad.

The final validated dataset for Data Release 1.0 consisted of 3,099 angiosperm samples (Table 1), only 5% fewer than were present in the initial dataset. These samples

represent all 64 orders, 404 families (96%; 212 represented by >1 sample), 2,333 genera (17%) and 2,956 species (0.01%).

Data Release 1.0: sequence quality and gene recovery

Nine statistics were used to assess the sequence quality across the 3,099 samples of Data Release 1.0 (Table 3). For the 2,374 target sequence capture samples, the mean percentage of on-target reads was 8%, the mean read depth per sample across all recovered genes was 90x with a median value of 38x and the mean percentage length of recovered genes per sample was 62%. The number of genes and the sum length of gene sequence recovered per sample were tightly correlated as expected, varying continuously across the dataset up to the full set of Angiosperms353 genes and a total gene length of 256.9 kbp, close to the maximum expected length of 260 kbp for recovering genes with this target gene set (Fig. 6). However, both the number of genes and sum length of gene sequence recovered were correlated less closely with the number of available reads than they were to each other. The total length of sequence recovered from target sequence capture data was shorter than for samples mined for Angiosperms353 genes from 1KP transcriptomes or annotated genomes data (Table 3). The reason for the shorter length of the recovered genes is that some exons were absent from the original 1KP alignments used by Johnson et al. (2019) to create the Angiosperms353 gene set. These missing exons are however present in 1KP transcriptomes and annotated genomes and were recovered during data mining. The variation in performance of target enrichment across different samples, illustrated by the measures of variability shown in Table 3, likely reflects the variation in structure and metabolite composition of the starting tissue, which is known to impede DNA extraction from various species and its downstream manipulation. This variation is one of the challenges in dealing with samples from a broad taxonomic range such as across the evolutionary diversity of angiosperms. Variation in gene recovery across orders is visualised in Supplementary Figure S3.

Phylogenetic results

The final phylogenetic tree as inferred from Data Release 1.0 is publicly available in interactive form via the Kew Tree of Life Explorer. In the current release, the tree is annotated with local posterior probabilities (LPP, as given by ASTRAL-III) as indicators of branch support. Other measures of support (e.g. quartet scores) can be found within tree files accessible via the RBGK secure FTP. For completeness, the tree is also available in various formats, including Newick (Supplementary Fig. S2).

As a result of filtering and trimming steps during alignment, six genes in Data Release 1.0 were excluded from downstream phylogenetic analysis due to insufficient overlap between sequences. All statistics provided below refer to the remaining dataset. Thus, the species tree is based on 347 gene alignments totalling 824,878 sequences, 489,086,049 base pairs and 532,260 alignment columns. Of these, 509,987 columns (96%) are variable and 475,181 columns (89%) are parsimony informative. The proportion of missing data across all alignments is 61.6% and the median number of genes per sample is 284 (mean: 265.3, standard deviation (SD): 64.3, min: 22, max: 347; Supplementary Table S2). The median number of samples per gene alignment is 2,421 (mean: 2,377.2, SD: 359) and median alignment length is 1,259 (mean: 1,533.9, SD: 985.7; Table 4). The resulting gene trees are highly resolved, with a median support across all nodes (ultrafast bootstrap) of 98% (mean: 87.8%, standard deviation (SD): 18.560) across all nodes in all gene trees (Fig. 7). Only 1.3% of all nodes in all gene trees are very poorly supported (ultrafast bootstrap <30%; Fig. 7) and thus collapsed prior to species tree inference. Further statistics for individual gene alignments and gene trees are reported in Table 4 and Supplementary Table S2.

The species tree accommodates 82% of the quartet relationships in the gene trees (ASTRAL normalized quartet score of 0.82). The majority (76.8%) of nodes in the species tree were well-supported (LPP ≥95%, cf. Sayyari and Mirarab 2016), and only seven nodes

were informed by too few gene trees (i.e. <20) to evaluate support. Comparing node support in the species tree at different taxonomic levels (Fig. 8), median quartet support is progressively higher towards shallower taxonomic levels (Fig. 8c), while the effective number of gene trees informing nodes shows the opposite trend (Fig. 8e). Local posterior probabilities show a tendency to be lower (1st quartile) at the deepest taxonomic level (Fig. 8a). Major groups (i.e. monocots, asterids and rosids) show similar distributions of both local posterior probabilities (Fig. 8b) and quartet support values (Fig. 8d), despite the fact that the effective number of gene trees supporting nodes is more variable in monocots (Fig. 8f), which is the result of the lower recovery rates for some orders in this group such as Alismatales, Commelinales and Liliales (Supplementary Fig. S3).

Discounting taxa represented by a single sample (193 families, one order), 96% of testable families and 83% of testable orders were resolved as monophyletic in the species tree. Most of the samples of non-monophyletic families and orders could be assigned to a clade that represents the family or order well, despite lacking some samples and/or containing some outlier samples from other taxa ("concordant taxa" where taxon concordance score >0.5, see Materials and Methods for details). Only five families (Francoaceae, Hernandiaceae, Phyllanthaceae, Pontederiaceae and Schlegeliaceae, represented by 11 samples) and two orders (Bruniales and Icacinales, represented by six samples) were so dispersed that this was not possible ("discordant taxa" where taxon concordance score ≤ 0.5). At the family level, 2,893 samples were resolved in the expected family, two samples were resolved in an unexpected position, and 204 samples were not testable because they belonged to a discordant family or a family represented by a single sample. At the order level, 3,060 samples were resolved in the expected order, 32 samples were resolved in an unexpected position, and seven samples were not testable (see Supplementary Tables S3-S5 for lists of specimens from singly represented taxa, poorly resolved taxa, and outliers to well-resolved

taxa, respectively). Placements of all but five genera and seven families were consistent with the WCVP/APG IV taxonomic hierarchy of genera, families and orders. Concordance with existing taxonomy was lower at the genus level, with only 74% of testable genera resolving as monophyletic and 47 genera (represented by 130 samples) being discordant; these numbers partly reflect the deliberate inclusion of multiple samples from genera suspected a priori to be potentially non-monophyletic.

In addition to resolving most genera, families and orders as monophyletic, our tree supports more than half (58%) of the relationships among orders presented by the Angiosperm Phylogeny Group (APG IV 2016; Supplementary Fig. S4). Congruence with APG IV varies among major clades, being notably high in magnoliids (100% of APG IV relationships supported) and monocots (80%), while being substantially lower in eudicots (47%), especially in rosids (33%). Nodes in our tree that are congruent with APG IV ordinal relationships are slightly better supported on average (mean LPP 0.98, median 1) than nodes that are incongruent with APG IV (mean LPP 0.75, median 0.94).

Tree of Life Explorer

The Kew Tree of Life Explorer (https://treeoflife.kew.org) provides open access to taxon, specimen, sequence, alignment and tree data, with associated metadata for the current data release in accordance with the Toronto guidelines on pre-publication data sharing (Toronto International Data Release Workshop Authors 2009). Users can browse by species, gene or interactive phylogenetic tree. The species interface permits searches by order, family, genus or species, and provides voucher specimen metadata (including links to online specimen images, where available), simple sequence metrics, access to assembled genes and raw data. The gene interface documents all Angiosperms353 genes and associated metrics, links to gene identities in UniProt (https://www.uniprot.org/) and provides access to assembled genes across taxa. The tree of life interface enables browsing and taxon searching

591 of the species tree inferred from the current release dataset, as well as tree downloads (as
592 PNG or Newick) and zooming into user-defined subtrees. All processed data (assembled
593 genes, alignments, gene trees, species trees) and archived releases are available from
594 RBGK's secure FTP site (http://sftp.kew.org/pub/treeoflife/), whereas raw sequence reads are
595 deposited within the European Nucleotide Archive (project number PRJEB35285) for
596 integration within the Sequence Read Archive.

597 **DISCUSSION**

598

599 The new phylogenomic platform described here is a major milestone towards a
600 comprehensive tree of life for all flowering plant species. Firstly, the sequencing of a
601 standardised nuclear marker set of this scale for so many taxa is unprecedented, opening
602 doors to a highly integrated future for angiosperm phylogenetics in the genomic era. Much
603 like a "next generation" *rbcL*, which underpinned so many Sanger sequencing-based plant
604 phylogenetic studies, the Angiosperms353 genes offer opportunities for continuous synthesis
605 of HTS data across angiosperms. The foundational dataset presented here can be re-used or
606 extended for tree of life research at almost any taxonomic scale (Johnson et al. 2019;
607 Larridon et al. 2019; Van Andel et al. 2019; Murphy et al. 2020; Pérez-Escobar et al. 2020;
608 Shee et al. 2020; Slimp et al. 2020; Beck et al. 2021). Secondly, this is the first phylogenetic
609 project to gather novel HTS data across angiosperms with a stratified taxon sampling at the
610 genus level. Our sampling strategy systematically and comprehensively represents both the
611 diversity of angiosperms and their deep-time diversification. As genus-level sampling
612 becomes increasingly complete—a target that is well within reach—this backbone will
613 substantially increase our ability to study the dynamics of plant diversity over time and revisit
614 long-standing questions in systematics (Magallón et al. 2018; Sauquet and Magallón 2018;

615 Soltis et al. 2019). Importantly, it will also sharpen the focus on truly intractable phylogenetic
616 problems (Yang et al. 2020; Zhao et al. 2020), encouraging the exploration of the biological
617 drivers of these phenomena.

618 Our approach has already led to a burst of community engagement. More than a
619 dozen studies utilising Angiosperms353 probes are already published (e.g. Larridon et al.
620 2019; Howard et al. 2020; Murphy et al. 2020; Pérez-Escobar et al. 2020; Shee et al. 2020;
621 Slimp et al. 2020; McLay et al. in press), and two journal special issues focused on the probe
622 set are in preparation arising from a recent symposium (Lagomarsino and Jabaily 2020). The
623 probe set has also been adopted by the Genomics for Australian Plants consortium
624 (https://www.genomicsforaustralianplants.com/), which aims to sequence all Australian
625 angiosperm genera, coordinating with the PAFTOL project to optimise collective taxonomic
626 coverage. A subset of the Angiosperms353 genes is now accessible for non-angiosperm land
627 plants thanks to a probe set developed in parallel (Breinholt et al. 2021), inviting the prospect
628 of data integration across all land plants. Angiosperms353 genes (as distinct from the
629 Angiosperms353 probes) are also being leveraged as components of custom-designed probe
630 sets (e.g. Jantzen et al. 2020; Ogutcen et al. 2021). This approach gives all the integrative
631 benefits of Angiosperms353, while permitting (i) the tailoring of Angiosperms353 probes to a
632 specific taxonomic group to increase gene recovery, and (ii) the inclusion of additional loci
633 pertinent to the research in question. Angiosperms353 probes have also been directly
634 combined with an existing custom probe set (Nikolov et al. 2019) as a "probe cocktail" in a
635 single hybridisation, capturing both sets of targets simultaneously with remarkable efficiency
636 (Hendriks et al. in press). These possibilities render the invidious choice between specific and
637 universal probe sets increasingly irrelevant (Kadlec et al. 2017).

638 We took several open data measures to encourage community uptake, in both the
639 design of our tools and the sharing of our data. The Angiosperms353 probe set itself was

designed to be a transparent, "off-the-shelf" toolkit that is open, inexpensive and accessible to all, especially researchers discouraged by the complexity and cost of custom probe design (Johnson et al. 2019). Our sequence data for Angiosperms353 genes are openly available via the Kew Tree of Life Explorer and the Sequence Read Archive, as a public foundation dataset shared according to pre-publication best practice (Toronto International Data Release Workshop Authors 2009). The Explorer offers enhanced transparency and accessibility by allowing users to navigate the data via a phylogenetic snapshot of the current release, along with metadata (e.g. specimen data) and intermediate data (e.g. gene assemblies, alignments, gene trees). Thanks to these resources, cross-community collaboration via Angiosperms353 is gaining momentum.

Our tree, which is based on the most extensive nuclear phylogenomic dataset in flowering plants to date, is strongly supported, credible and highly congruent with existing taxonomy and many hypothesized relationships among orders (APG IV 2016; Supplementary Fig. S4). The data confirm both the effectiveness of Angiosperms353 probes across all major angiosperm clades and the ability of the genes to resolve relationships across taxonomic scales (Fig. 8). Variable sequence recovery notwithstanding (Table 3, Supplementary Fig. S3), most nodes in our tree are underpinned by large numbers of gene trees (Fig. 8e), allowing the species tree to be inferred with confidence (Fig. 8a) despite gene tree conflict (Fig. 8c). However, even the most strongly supported phylogenetic hypotheses must be viewed with caution as they may be biased by model misspecification and wrong assumptions. Moreover, our "first pass" analyses based on a set of standard methods may not suit this dataset perfectly (see below). Nevertheless, our findings are rendered credible by their high concordance with taxonomy, an independent point of reference that has been extensively ground-truthed by pre-phylogenomic DNA data, especially plastid loci. Agreement with existing family circumscriptions is particularly striking. In contrast,

665 congruence with previously hypothesized relationships among orders (APG IV 2016) is much
666 lower (Supplementary Fig. S4). Some of these earlier hypothesized ordinal relationships
667 derive from relatively weak evidence (bootstrap/jackknife >50%; APG IV 2016), which may
668 partly explain this disagreement. However, it may also be due to phylogenetic conflict
669 between nuclear and plastid genomes, as the established ordinal relationships rest primarily
670 on evidence from plastid loci, substantiated more recently by plastid genomes (Li et al.
671 2019). It is hardly surprising, then, that a large-scale nuclear analysis presents strongly
672 supported, alternative relationships (Supplementary Fig. S4). The conundrum remains that
673 these incongruences are visible at the ordinal backbone, but not the family level. A more
674 comprehensive exploration of these relationships, the underlying phylogenetic signal and
675 their systematic implications is currently underway.

676 The analyses presented here are primarily intended as a window onto the information
677 content of our current data release and are not a complete exploration of the data. Thus,
678 downstream application of the current species tree comes with caveats. We used current,
679 widely accepted methods in a pipeline that can be re-run in a semi-automated fashion
680 whenever we release new data. As a consequence, not all possible analysis options and
681 effects have been explored. We anticipate that users of our data will probe it more rigorously
682 and will tailor both sampling and phylogenomic analyses to their specific questions.

683 Important limitations in our analysis relate to (i) sampling, (ii) gene recovery, (iii)
684 models of sequence evolution and (iv) paralogy. Sampling for intermediate data releases is
685 biased by the current state of progress towards our systematic sampling strategy. This will be
686 addressed in future data releases and can be adjusted by users of our data. Gene recovery
687 relied upon the standard Angiosperms353 target file (Johnson et al. 2019), but it has recently
688 become apparent that tailoring target sequences to taxonomic groups can improve recovery
689 (McLay et al. in press); this will be tested in future releases. Moreover, we are yet to exploit

intronic data captured in the "splash zone" adjacent to our target exons. By necessity, our "first pass" phylogenetic analysis does not explore the fast-evolving spectrum of methodological options available for phylogenomic analysis. For example, we rely on a simple standard model of sequence evolution, but more sophisticated models accounting for codon positions or amino acids may improve phylogenetic inference. Potential paralogy is not addressed by our current pipeline. The genes underpinning our analysis were carefully chosen to represent single-copy genes across flowering plants (Johnson et al. 2019; Leebens-Mack et al. 2019). However, some paralogy may have gone unnoticed due to the pervasiveness of gene and genome duplication in plants (Li and Barker 2020). Overall, we expect that the occasional presence of paralogs in our current analysis would more likely lead to inflated estimates of gene tree incongruence, and thus result in reduced support values, than significant topological biases (Yan et al. 2020). Thus, we consider our tree relatively conservative while acknowledging that we are not yet exploiting the full potential of our data. Although a rigorous analysis of paralogy in Angiosperms353 genes was not tractable for this data release, we look forward to deeper insights emerging as community-wide engagement with Angiosperms353 grows.

PROSPECTS

In the immediate future, we will deliver a further data release through which we expect to reach the milestone of sampling 50% of all angiosperm genera. This target will be achieved through substantial novel data production by PAFTOL and collaborators, augmented by data mined from public sources. In-depth phylogenetic analyses of our data and their evolutionary implications are also underway.

Beyond this point, we see three priority areas in which future platform developments might be concentrated, resources permitting. Firstly, taxon sampling to the genus level must be completed. Our original target of sampling all angiosperm genera remains, but the mode of reaching this is likely to evolve. We anticipate an acceleration in production of Angiosperms353 data by the broader community. The completion of generic-level sampling will require both the integration of community data in the broader angiosperm tree of life as well as strategic investment in filling inevitable data gaps for orphan groups. Secondly, numerous opportunities for refinement exist across our methods. For example, insights from our data might permit the optimisation of the Angiosperms353 probes to improve gene capture. Efficiency of gene assembly from sequence data can also be improved bioinformatically (McLay et al. in press). As costs of sequencing decline, target sequence capture *in vitro* may no longer be necessary, the target genes being retrieved simply from sufficiently deeply sequenced genomes. Thirdly, for the full integrative potential of Angiosperms353 genes to be achieved, infrastructure for aggregating and sharing this coherent body of data must be improved. While the Kew Tree of Life Explorer provides a proof-of-concept, it is the public data repositories (e.g. NCBI, ENA) that offer the greatest prospects of a mechanism to achieve this. To fully parallel the earlier success of public repositories for facilitating single-gene phylogenetic trees (e.g. *rbcL*, *matK*), new tools are needed to assist with efficient upload and annotation of target capture loci and associated metadata.

Even with a completed genus-level angiosperm tree of life well within reach, the monumental task of sampling all species remains. The scale of this challenge is 24-fold greater than the genus-level tree towards which we are currently working. However, with sufficient investment, increased efficiencies and community engagement, such an ambition could potentially be realised. Collections-based institutions are poised to play a critical role in

this endeavour through increasingly routine molecular characterisation of their specimens, perhaps as part of digitisation programmes, and are already facilitating the growing trend towards species-complete sampling in phylogenomic studies (e.g. Loiseau et al. 2019; Murphy et al. 2020; Kuhnhäuser et al. 2021). Our platform demonstrates how large-scale phylogenomic projects can capitalise on natural history collections to achieve a much more complete sampling than hitherto possible.

The growing movement to sequence the genomes of all life on Earth, inspired by the Earth Biogenome Project (Lewin et al. 2018), significantly boosts the prospects for completing the tree of life for all species, but is hampered by the focus on "gold standard" whole genomes requiring the highest quality input DNA. Our platform offers the opportunity to bridge the gap between the ambition of these projects and the vast phylogenomic potential of natural history collections. However, as life on Earth becomes increasingly imperilled, we cannot afford to wait. To meet the urgent demand for best estimates of the tree of life, we must dynamically integrate phylogenetic information as it is generated, providing synthetic trees of life to the broadest community of potential users (Eiserhardt et al. 2018). Our platform facilitates this crucial synthesis by providing a cross-cutting dataset and directing the community towards universal markers that seem set to play a central role in completing an integrated angiosperm tree of life.

DATA AVAILABILITY AND SUPPLEMENTARY MATERIAL

All data generated in this study are publicly released under a Creative Commons Attribution 4.0 International (CC BY 4.0) license and the Toronto guidelines on pre-publication data sharing (Toronto International Data Release Workshop Authors 2009). The data are accessible via the Kew Tree of Life Explorer (https://treeoflife.kew.org) and our

763 secure FTP (http://sftp.kew.org/pub/treeoflife/). Raw sequence reads are deposited in the
764 European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) under umbrella
765 project PRJEB35285. Scripts and other files relating to our phylogenomic pipeline are
766 available at our GitHub (https://github.com/RBGKew/KewTreeOfLife). Supplementary
767 materials cited in this paper are available from the Dryad Digital Repository
768 (http://dx.doi.org/10.5061/dryad.[NNNN]).

769 **FUNDING**

770

771 This work was supported by grants from the Calleva Foundation and the Sackler Trust
772 to the Plant and Fungal Trees of Life project at the Royal Botanic Gardens, Kew. Additional
773 support was received from the Garfield Weston Foundation, as part of the Global Tree Seed
774 Bank Programme.

775 **ACKNOWLEDGEMENTS**

776 We would like to thank Guilherme Antar, Alex Antonelli, Marc Appelhans, Julien
777 Bachelier, Donovan Bailey, Aurélien Bour, Peter Boyce, Gemma Bramley, Sven Buerki,
778 Stuart Cable, Martin Callmander, Monica Carlsen, Vinicius Castro Sousa, Mark Chase,
779 Martin Cheek, Maarten Christenhusz, Thomas Couvreur, Darren Crayn, Iain Darbyshire,
780 Alison Devault, Manuel de la Estrella, Elton John de Lirio, Jurriaan de Vos, Zacky Ezedin,
781 Federico Fabriani, Mike Fay, Geneviève Ferry, Helen Fortune-Hopkins, Jocelyn Hall, Ameka
782 Gabriel Komla, Jim Leebens-Mack, Elliot Gardner, Ester Gaya, Mark Gibernau, Olwen
783 Grace, Sean Graham, Jan Hackel, Anna Haigh, Kasper Hendriks, Oriane Hidalgo, Elizabeth
784 Joyce, Bente Klitgaard, Sophie Lane, Isabel Larridon, Drew Larson, Frederic Lens, Christine
785 Leon, Gwil Lewis, Jing-Xia Liu, Meng Lu, Jaquelini Luber, Eve Lucas, Penny Malakasi,

Vidal Mansano, Laura Martinez-Suz, Angela McDonell, Alexander Monro, Michael Moore, Klaus Mummenhof, Tuula Niskanen, Andres Orejuela, Luis Palazzesi, Joe Parker, Frederic Pautz, Jaume Pellicer, Oscar Perez Escobar, Yohan Pillon, Jose Pirani, Robyn Powell, Natalia Przelomska, Carmen Puglisi, Eric Roalson, Hervé Sauquet, Hanno Schaefer, Ruud Scharn, Rowan Schley, David Scherberich, Toral Shah, Mark P. Simmons, Ana Rita Simões, Lalita Simpson, Stephen Smith, Doug Soltis, Pam Soltis, Cynthia Sothers, Marybel Soto Gomez, Jemma Taylor, Liam Trethowan, Anna Trias-Blasi, Tim Utteridge, Juan Viruel, Maria Vorontsova, Gane Ka-Shu Wong, Sin Yeng Wong and Sue Zmarzty for helping PAFTOL reach its goals through collaboration, sharing expertise and providing samples; Noelia Alvarez de Roman, Richard Barley, Nicola Biggs, Elisa Biondi, Elinor Breman, Hannah Button, Christopher Cockel, David Cooke, Nina Davies, Solene Dequiret, John Dickie, Florence Ducan-Antoine, Sara Edwards, Thomas Freeth, Sue Frisby, Tim Fulcher, Aurélie Grall, Anthony Hall, Alex Hankey, Kate Hardwick, Keegan Hickey, David Hickmott, Rebecca Hilgenhof, Imalka Kahandawala, Lara Jewitt, Laura Jennings, Nick Johnson, Udayangani Liu, Carlos Magdalena, Max Moog, Richard Moore, Ana Oliveira, Tim Pearce, Tom Pickering, Sara Redstone, Greg Redwood, Luxy Reed, Paul Rees, Matthew Rees, Silke Roch, Daniel Rosenberg, Marcello Sellaro, Scott Taylor, Janet Terry, Michael Way, Ian Willey, Patricia Woods, Rosie Woods and Martin Xanthos for support with acquisition samples from RBGK collections, both living and preserved; Alexander Bowles, Dion Devey, Laszlo Csiba, Isabel Fairlie, Lorna Frankel, Karime Gutierrez, Alina Höwener, Izai A. B. Sabino Kikuchi, Beata Klejevskaja, Jake Newitt, Michelle Siros and Jessica Tengvall, Haydn Thompson, for assistance with laboratory work and data collection; Laura Green, Alan Paton, Sarah Phillips and Marie-Helene Weech for support with specimen digitisation; Nicholas Black, Michael Bradford, Carol Sinker, Robert Turner and Noor Al Wattar for assistance

with computational infrastructure. Finally, special thanks to Kathy Willis, former Director of Science at RBGK, for inspiring the establishment of the PAFTOL project.

LITERATURE CITED

Abadi S., Azouri D., Pupko T., Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. Nat. Commun. 10:934.

Alsos I.G., Lavergne S., Merkel M.K., Boleda M., Lammers Y., Alberti A., Pouchon C., Denoeud F., Pitelkova I., Pușcaș M., Roquet C., Hurdu B.-I., Thuiller W., Zimmermann N.E., Hollingsworth P.M., Coissac E. 2020. The treasure vault can be opened: Large-scale genome skimming works well using herbarium and silica gel dried material. Plants 9:432.

Antonelli A., Fry C., Smith R.J., Simmonds M.S.J., Kersey P.J., Pritchard H.W., Abbo M.S., Acedo C., Adams J., Ainsworth A.M., Allkin B., Annecke W., Bachman S.P., Bacon K., Bárrios S., Barstow C., Battison A., Bell E., Bensusan K., Bidartondo M.I., Blackhall-Miles R.J., Borrell J.S., Brearley F.Q., Breman E., Brewer R.F.A., Brodie J., Cámara-Leret R., Campostrini Forzza R., Cannon P., Carine M., Carretero J., Cavagnaro T.R., Cazar M.E., Chapman T., Cheek M., Clubbe C., Cockel C., Collemare J., Cooper A., Copeland A.I., Corcoran M., Couch C., Cowell C., Crous P., da Silva M., Dalle G., Das D., David J.C., Davies L., Davies N., De Canha M.N., de Lirio E.J., Demissew S., Diazgranados M., Dickie J., Dines T., Douglas B., Dröge G., Dulloo M.E., Fang R., Farlow A., Farrar K., Fay M.F., Felix J., Forest F., Forrest L.L., Fulcher T., Gafforov Y., Gardiner L.M., Gâteblé G., Gaya E., Geslin B., Gonçalves S.C., Gore C.J.N., Govaerts R., Gowda B., Grace O.M., Grall A., Haelewaters D., Halley J.M., Hamilton M.A., Hazra A., Heller T., Hollingsworth P.M., Holstein N., Howes M.J.R., Hughes M., Hunter D., Hutchinson N., Hyde K., Iganci J., Jones

835 M., Kelly L.J., Kirk P., Koch H., Grisai-Greilhuber I., Lall N., Langat M.K., Leaman D.J.,
836 Leão T.C., Lee M.A., Leitch I.J., Leon C., Lettice E., Lewis G.P., Li L., Lindon H., Liu J.S.,
837 Liu U., Llewellyn T., Looney B., Lovett J.C., Luczaj L., Lulekal E., Maggassouba S.,
838 Malécot V., Martin C., Masera O.R., Mattana E., Maxted N., Mba C., McGinn K.J.,
839 Metheringham C., Miles S., Miller J., Milliken W., Moat J., Moore P.G.P., Morim M.P.,
840 Mueller G.M., Muminjanov H., Negrão R., Nic Lughadha E., Nicholson N., Niskanen T.,
841 Nono Womdim R., Noorani A., Obreza M., O'Donnell K., O'Hanlon R., Onana J.M., Ondo I.,
842 Padulosi S., Paton A., Pearce T., Pérez Escobar O.A., Pieroni A., Pironon S., Prescott T.A.K.,
843 Qi Y.D., Qin H., Quave C.L., Rajaovelona L., Razanajatovo H., Reich P.B., Rianawati E.,
844 Rich T.C.G., Richards S.L., Rivers M.C., Ross A., Rumsey F., Ryan M., Ryan P., Sagala S.,
845 Sanchez M.D., Sharrock S., Shrestha K.K., Sim J., Sirakaya A., Sjöman H., Smidt E.C.,
846 Smith D., Smith P., Smith S.R., Sofo A., Spence N., Stanworth A., Stara K., Stevenson P.C.,
847 Stroh P., Suz L.M., Tambam B.B., Tatsis E.C., Taylor I., Thiers B., Thormann I., Vaglica V.,
848 Vásquez-Londoño C., Victor J., Viruel J., Walker B.E., Walker K., Walsh A., Way M.,
849 Wilbraham J., Wilkin P., Wilkinson T., Williams C., Winterton D., Wong K.M., Woodfield-
850 Pascoe N., Woodman J., Wyatt L., Wynberg R., Zhang B.G. 2020. State of the World's Plants
851 and Fungi 2020. Royal Botanic Gardens, Kew.
852
853 APG. 1998. An ordinal classification for the families of flowering plants. Ann. Missouri Bot.
854 Gard. 85:531-553.
855
856 APG II. 2003. An update of the Angiosperm Phylogeny Group classification for the orders
857 and families of flowering plants: Apg II. Bot. J. Linn. Soc. 141:399-436.
858

APG III. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: Apg III. Bot. J. Linn. Soc. 161:105-121.

APG IV. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: Apg IV. Bot. J. Linn. Soc. 181:1-20.

Bakker F.T., Lei D., Yu J., Mohammadin S., Wei Z., van de Kerke S., Gravendeel B., Nieuwenhuis M., Staats M., Alquezar-Planas D.E., Holmer R. 2016. Herbarium genomics: Plastome sequence assembly from a range of herbarium specimens using an iterative organelle genome assembly pipeline. Biol. J. Linn. Soc. 117:33-43.

Beck J.B., Markley M.L., Zielke M.G., Thomas J.R., Hale H.J., Williams L.D., Johnson M.G. 2021. Is Palmer's elm leaf goldenrod real? The Angiosperms353 kit provides within-species signal in *Solidago ulmifolia* s.L. bioRxiv:2021.2001.2007.425781.

Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Borowiec M.L. 2016. AMAS: A fast tool for alignment manipulation and computing of summary statistics. PeerJ 4:e1660.

Bostock M. 2012. D3.Js - data-driven documents http://d3js.Org/.

Breinholt J.W., Carey S.B., Tiley G.P., Davis E.C., Endara L., McDaniel S.F., Neves L.G., Sessa E.B., von Konrat M., Chantanaorrapint S., Fawcett S., Ickert-Bond S.M., Labiak P.H.,

884 Larraín J., Lehnert M., Lewis L.R., Nagalingum N.S., Patel N., Rensing S.A., Testo W.,
885 Vasco A., Villarreal J.C., Williams E.W., Burleigh J.G. 2021. A target enrichment probe set
886 for resolving the flagellate land plant tree of life. Appl. Plant. Sci. n/a:e11406.
887
888 Brewer G.E., Clarkson J.J., Maurin O., Zuntini A.R., Barber V., Bellot S., Biggs N., Cowan
889 R.S., Davies N.M.J., Dodsworth S., Edwards S.L., Eiserhardt W.L., Epitawalage N., Frisby
890 S., Grall A., Kersey P.J., Pokorny L., Leitch I.J., Forest F., Baker W.J. 2019. Factors
891 affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the
892 diversity of angiosperms. Front. Plant Sci. 10:1102.
893
894 Buddenhagen C., Lemmon A.R., Lemmon E.M., Bruhl J., Cappa J., Clement W.L.,
895 Donoghue M.J., Edwards E.J., Hipp A.L., Kortyna M. 2016. Anchored phylogenomics of
896 angiosperms I: Assessing the robustness of phylogenetic estimates. bioRxiv:086298.
897
898 Buerki S., Baker W.J. 2016. Collections-based research in the genomic era. Biol. J. Linn.
899 Soc. 117:5-10.
900
901 Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L.
902 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10:421.
903
904 Carpenter E.J., Matasci N., Ayyampalayam S., Wu S., Sun J., Yu J., Jimenez Vieira F.R.,
905 Bowler C., Dorrell R.G., Gitzendanner M.A., Li L., Du W., K. Ullrich K., Wickett N.J.,
906 Barkmann T.J., Barker M.S., Leebens-Mack J.H., Wong G.K.-S. 2019. Access to rna-
907 sequencing data from 1,173 plant species: The 1000 Plant Transcriptomes Initiative (1KP).
908 GigaScience 8:giz126.

Chase M.W., Hills H.H. 1991. Silica gel: An ideal material for field preservation of leaf samples for DNA studies. Taxon 40:215-220.

Chase M.W., Soltis D.E., Olmstead R.G., Morgan D., Les D.H., Mishler B.D., Duvall M.R., Price R.A., Hills H.G., Qiu Y.L., Kron K.A., Rettig J.H., Conti E., Palmer J.D., Manhart J.R., Sytsma K.J., Michaels H.J., Kress W.J., Karol K.G., Clark W.D., Hedren M., Gaut B.S., Jansen R.K., Kim K.J., Wimpee C.F., Smith J.F., Furnier G.R., Strauss S.H., Xiang Q.Y., Plunkett G.M., Soltis P.S., Swensen S.M., Williams S.E., Gadek P.A., Quinn C.J., Eguiarte L.E., Golenberg E., Learn G.H., Graham S.W., Barrett S.C.H., Dayanandan S., Albert V.A. 1993. Phylogenetics of seed plants - an analysis of nucleotide sequences from the plastid gene *rbcL*. Ann. Missouri Bot. Gard. 80:528-580.

Chau J.H., Rahfeldt W.A., Olmstead R.G. 2018. Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. Appl. Plant. Sci. 6:e1032.

Cheng S., Melkonian M., Smith S.A., Brockington S., Archibald J.M., Delaux P.-M., Li F.-W., Melkonian B., Mavrodiev E.V., Sun W., Fu Y., Yang H., Soltis D.E., Graham S.W., Soltis P.S., Liu X., Xu X., Wong G.K.-S. 2018. 10kp: A phylodiverse genome sequencing plan. GigaScience 7:giy013.

Cornwell W.K., Pearse W.D., Dalrymple R.L., Zanne A.E. 2019. What we (don't) know about global plant diversity. Ecography 42:1819-1831.

933 Couvreur T.L.P., Helmstetter A.J., Koenen E.J.M., Bethune K., Brandão R.D., Little S.A.,
934 Sauquet H., Erkens R.H.J. 2019. Phylogenomics of the major tropical plant family
935 Annonaceae using targeted enrichment of nuclear genes. Front. Plant Sci. 9:1941.
936
937 Dodsworth S., Pokorny L., Johnson M.G., Kim J.T., Maurin O., Wickett N.J., Forest F.,
938 Baker W.J. 2019. Hyb-Seq for flowering plant systematics. Trends Plant Sci. 24:887-891.
939
940 Doyle J.J., Doyle J.L. 1987. A rapid DNA isolation procedure from small quantities of fresh
941 leaf tissue. Phytochem. Bull. 19:11-15.
942
943 Eiserhardt W.L., Antonelli A., Bennett D.J., Botigué L.R., Burleigh J.G., Dodsworth S.,
944 Enquist B.J., Forest F., Kim J.T., Kozlov A.M., Leitch I.J., Maitner B.S., Mirarab S., Piel
945 W.H., Pérez-Escobar O.A., Pokorny L., Rahbek C., Sandel B., Smith S.A., Stamatakis A.,
946 Vos R.A., Warnow T., Baker W.J. 2018. A roadmap for global synthesis of the plant tree of
947 life. Amer. J. Bot. 105:614-622.
948
949 Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C.
950 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple
951 evolutionary timescales. Syst. Biol. 61:717-726.
952
953 Forrest L.L., Hart M.L., Hughes M., Wilson H.P., Chung K.-F., Tseng Y.-H., Kidner C.A.
954 2019. The limits of Hyb-Seq for herbarium specimens: Impact of preservation techniques.
955 Front. Ecol. Evol. 7:439.
956

957 Gitzendanner M.A., Soltis P.S., Wong G.K.-S., Ruhfel B.R., Soltis D.E. 2018. Plastid
958 phylogenomic analysis of green plants: A billion years of evolutionary history. Amer. J. Bot.
959 105:291-301.
960
961 Hale H., Gardner E.M., Viruel J., Pokorny L., Johnson M.G. 2020. Strategies for reducing
962 per-sample costs in target capture sequencing for phylogenomics and population genomics in
963 plants. Appl. Plant. Sci. 8:e11337.
964
965 Hendriks K., Mandáková T., Hay N.M., Ly E., Hooft van Huysduynen A., Tamrakar R.,
966 Thomas S.K., Toro-Núñez O., Pires J.C., Nikolov L.A., Koch M.A., Windham M.D., Lysak
967 M.A., Forest F., Mummenhoff K., Baker W.J., Lens F., Bailey C.D. in press. The best of both
968 worlds: Combining lineage specific and universal bait sets in target enrichment hybridization
969 reactions. Appl. Plant. Sci.
970
971 Hinchliff C.E., Smith S.A. 2014. Some limitations of public sequence data for phylogenetic
972 inference (in plants). PLoS ONE 9:e98986.
973
974 Hinchliff C.E., Smith S.A., Allman J.F., Burleigh J.G., Chaudhary R., Coghill L.M., Crandall
975 K.A., Deng J., Drew B.T., Gazis R., Gude K., Hibbett D.S., Katz L.A., Laughinghouse H.D.,
976 McTavish E.J., Midford P.E., Owen C.L., Ree R.H., Rees J.A., Soltis D.E., Williams T.,
977 Cranston K.A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life.
978 Proc. Natl. Acad. Sci. U.S.A. 112:12764.
979
980 Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2017. UFBoot2:
981 Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518-522.

Howard C.C., Crowl A.A., Harvey T.S., Cellinese N. 2020. Peeling back the layers: The complex dynamics shaping the evolution of the Ledebouriinae (Scilloideae, Asparagaceae). bioRxiv:2020.2011.2002.365718.

Jantzen J.R., Amarasinghe P., Folk R.A., Reginato M., Michelangeli F.A., Soltis D.E., Cellinese N., Soltis P.S. 2020. A two-tier bioinformatic pipeline to develop probes for target capture of nuclear loci with applications in Melastomataceae. Appl. Plant. Sci. 8:e11345.

Jin J.-J., Yu W.-B., Yang J.-B., Song Y., dePamphilis C.W., Yi T.-S., Li D.-Z. 2020. GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 21:241.

Johnson M.G., Pokorny L., Dodsworth S., Botigue L.R., Cowan R.S., Devault A., Eiserhardt W.L., Epitawalage N., Forest F., Kim J.T., Leebens-Mack J.H., Leitch I.J., Maurin O., Soltis D.E., Soltis P.S., Wong G.K., Baker W.J., Wickett N.J. 2019. 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.

Junier T., Zdobnov E.M. 2010. The newick utilities: High-throughput phylogenetic tree processing in the Unix shell. Bioinformatics 26:1669-1670.

Kadlec M., Bellstedt D.U., Le Maitre N.C., Pirie M.D. 2017. Targeted NGS for species level phylogenomics:“Made to measure” or “one size fits all”? PeerJ 5:e3569.

1007 Kreft Ł., Botzki A., Coppens F., Vandepoele K., Van Bel M. 2017. Phyd3: A phylogenetic
1008 tree viewer with extended phyloXML support for functional genomics data visualization.
1009 Bioinformatics 33:2946-2947.
1010
1011 Kuhnhäuser B.G., Bellot S., Couvreur T.L.P., Dransfield J., Henderson A., Schley R.,
1012 Chomicki G., Eiserhardt W.L., Hiscock S.J., Baker W.J. 2021. A robust phylogenomic
1013 framework for the calamoid palms. Mol. Phylogenet. Evol.:107067.
1014
1015 Lagomarsino L.P., Jabaily R.S. 2020. Virtual Botany Conference 2020 symposium -
1016 Angiosperms353: A new essential tool for plant systematics.
1017 http://2020.botanyconference.org/engine/search/index.php?func=detail&aid=941.
1018
1019 Larridon I., Villaverde T., Zuntini A.R., Pokorny L., Brewer G.E., Epitawalage N., Fairlie I.,
1020 Hahn M., Kim J., Maguilla E., Maurin O., Xanthos M., Hipp A.L., Forest F., Baker W.J.
1021 2019. Tackling rapid radiations with targeted sequencing. Front Plant Sci 10:1655.
1022
1023 Leebens-Mack J.H., Barker M.S., Carpenter E.J., Deyholos M.K., Gitzendanner M.A.,
1024 Graham S.W., Grosse I., Li Z., Melkonian M., Mirarab S., Porsch M., Quint M., Rensing
1025 S.A., Soltis D.E., Soltis P.S., Stevenson D.W., Ullrich K.K., Wickett N.J., DeGironimo L.,
1026 Edger P.P., Jordon-Thaden I.E., Joya S., Liu T., Melkonian B., Miles N.W., Pokorny L.,
1027 Quigley C., Thomas P., Villarreal J.C., Augustin M.M., Barrett M.D., Baucom R.S., Beerling
1028 D.J., Benstein R.M., Biffin E., Brockington S.F., Burge D.O., Burris J.N., Burris K.P.,
1029 Burtet-Sarramegna V., Caicedo A.L., Cannon S.B., Çebi Z., Chang Y., Chater C., Cheeseman
1030 J.M., Chen T., Clarke N.D., Clayton H., Covshoff S., Crandall-Stotler B.J., Cross H.,
1031 dePamphilis C.W., Der J.P., Determann R., Dickson R.C., Di Stilio V.S., Ellis S., Fast E.,

1032 Feja N., Field K.J., Filatov D.A., Finnegan P.M., Floyd S.K., Fogliani B., García N., Gâteblé
1033 G., Godden G.T., Goh F., Greiner S., Harkess A., Heaney J.M., Helliwell K.E., Heyduk K.,
1034 Hibberd J.M., Hodel R.G.J., Hollingsworth P.M., Johnson M.T.J., Jost R., Joyce B., Kapralov
1035 M.V., Kazamia E., Kellogg E.A., Koch M.A., Von Konrat M., Könyves K., Kutchan T.M.,
1036 Lam V., Larsson A., Leitch A.R., Lentz R., Li F.-W., Lowe A.J., Ludwig M., Manos P.S.,
1037 Mavrodiev E., McCormick M.K., McKain M., McLellan T., McNeal J.R., Miller R.E.,
1038 Nelson M.N., Peng Y., Ralph P., Real D., Riggins C.W., Ruhsam M., Sage R.F., Sakai A.K.,
1039 Scascitella M., Schilling E.E., Schlösser E.-M., Sederoff H., Servick S., Sessa E.B., Shaw
1040 A.J., Shaw S.W., Sigel E.M., Skema C., Smith A.G., Smithson A., Stewart C.N.,
1041 Stinchcombe J.R., Szövényi P., Tate J.A., Tiebel H., Trapnell D., Villegente M., Wang C.-N.,
1042 Weller S.G., Wenzel M., Weststrand S., Westwood J.H., Whigham D.F., Wu S., Wulff A.S.,
1043 Yang Y., Zhu D., Zhuang C., Zuidof J., Chase M.W., Pires J.C., Rothfels C.J., Yu J., Chen
1044 C., Chen L., Cheng S., Li J., Li R., Li X., Lu H., Ou Y., Sun X., Tan X., Tang J., Tian Z.,
1045 Wang F., Wang J., Wei X., Xu X., Yan Z., Yang F., Zhong X., Zhou F., Zhu Y., Zhang Y.,
1046 Ayyampalayam S., Barkman T.J., Nguyen N.-p., Matasci N., Nelson D.R., Sayyari E.,
1047 Wafula E.K., Walls R.L., Warnow T., An H., Arrigo N., Baniaga A.E., Galuska S., Jorgensen
1048 S.A., Kidder T.I., Kong H., Lu-Irving P., Marx H.E., Qi X., Reardon C.R., Sutherland B.L.,
1049 Tiley G.P., Welles S.R., Yu R., Zhan S., Gramzow L., Theißen G., Wong G.K.-S., One
1050 Thousand Plant Transcriptomes I. 2019. One thousand plant transcriptomes and
1051 the phylogenomics of green plants. Nature 574:679-685.
1052
1053 Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored hybrid enrichment for massively
1054 high-throughput phylogenomics. Syst. Biol. 61:727-744.
1055

Lewin H.A., Robinson G.E., Kress W.J., Baker W.J., Coddington J., Crandall K.A., Durbin R., Edwards S.V., Forest F., Gilbert M.T.P., Goldstein M.M., Grigoriev I.V., Hackett K.J., Haussler D., Jarvis E.D., Johnson W.E., Patrinos A., Richards S., Castilla-Rubio J.C., van Sluys M.-A., Soltis P.S., Xu X., Yang H., Zhang G. 2018. Earth Biogenome Project: Sequencing life for the future of life. Proc. Natl. Acad. Sci. U.S.A. 115:4325-4333.

Li H.-T., Yi T.-S., Gao L.-M., Ma P.-F., Zhang T., Yang J.-B., Gitzendanner M.A., Fritsch P.W., Cai J., Luo Y., Wang H., van der Bank M., Zhang S.-D., Wang Q.-F., Wang J., Zhang Z.-R., Fu C.-N., Yang J., Hollingsworth P.M., Chase M.W., Soltis D.E., Soltis P.S., Li D.-Z. 2019. Origin of angiosperms and the puzzle of the Jurassic gap. Nat. Plants 5:461-470.

Li Z., Barker M.S. 2020. Inferring putative ancient whole-genome duplications in the 1000 Plants (1KP) Initiative: Access to gene family phylogenies and age distributions. GigaScience 9:giaa004.

Loiseau O., Olivares I., Paris M., de La Harpe M., Weigand A., Koubinova D., Rolland J., Bacon C.D., Balslev H., Borchsenius F. 2019. Targeted capture of hundreds of nuclear genes unravels phylogenetic relationships of the diverse neotropical palm tribe Geonomateae. Front. Plant Sci. 10:864.

Magallón S., Sánchez-Reyes L.L., Gómez-Acevedo S.L. 2018. Thirty clues to the exceptional diversification of flowering plants. Ann. Bot. 123:491-503.

Mai U., Mirarab S. 2018. TreeShrink: Fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics 19:272.

Mandel J.R., Dikow R.B., Funk V.A., Masalia R.R., Staton S.E., Kozik A., Michelmore R.W., Rieseberg L.H., Burke J.M. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. Appl. Plant. Sci. 2:1300085.

McLay T.G.B., Gunn B.F., Ning W., Tate J.A., Nauheimer L., Joyce E.M., Simpson L., Schmidt-Lebuhn A.N., Baker W.J., Forest F., Jackson C.J. in press. New targets acquired: Improving locus recovery from the Angiosperms353 probe set. Appl. Plant. Sci.

Meyer M., Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols 2010:pdb.prot5448.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., Von Haeseler A., Lanfear R. 2020. Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37:1530-1534.

Murphy B., Forest F., Barraclough T., Rosindell J., Bellot S., Cowan R., Golos M., Jebb M., Cheek M. 2020. A phylogenomic analysis of *Nepenthes* (Nepenthaceae). Mol. Phylogenet. Evol. 144:106668.

Nguyen N.-P.D., Mirarab S., Kumar K., Warnow T. 2015. Ultra-large alignments using phylogeny-aware profiles. Genome Biol. 16:124.

Nikolov L.A., Shushkov P., Nevado B., Gan X., Al-Shehbaz I.A., Filatov D., Bailey C.D., Tsiantis M. 2019. Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. New Phytol. 222:1638-1651.

Ogutcen E., Christe C., Nishii K., Salamin N., Möller M., Perret M. 2021. Phylogenomics of Gesneriaceae using targeted capture of nuclear genes. Mol. Phylogenet. Evol.:107068.

Pérez-Escobar O.A., Dodsworth S., Bogarín D., Bellot S., Balbuena J.A., Schley R., Kikuchi I., Morris S.K., Epitawalage N., Cowan R., Maurin O., Zuntini A., Arias T., Serna A., Gravendeel B., Torres M.F., Nargar K., Chomicki G., Chase M.W., Leitch I.J., Forest F., Baker W.J. 2020. Hundreds of nuclear and plastid loci yield insights into orchid relationships. bioRxiv:2020.2011.2017.386508.

RBG Kew. 2015. A global resource for plant and fungal knowledge. Science strategy 2015-2020. Royal Botanic Gardens, Kew.

RBG Kew. 2016. The State of the World's Plants report – 2016. Royal Botanic Gardens, Kew.

Sauquet H., Magallón S. 2018. Key questions and challenges in angiosperm macroevolution. New Phytol. 219:1170-1187.

Sayyari E., Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33:1654-1668.

Secretariat of the Convention on Biological Diversity. 2011. Nagoya protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the convention on biological diversity. Montreal: United Nations Environment Programme.

Shee Z.Q., Frodin D.G., Cámara-Leret R., Pokorny L. 2020. Reconstructing the complex evolutionary history of the Papuasian *Schefflera* radiation through herbariomics. Front. Plant Sci. 11:258.

Slimp M., Williams L.D., Hale H., Johnson M.G. 2020. On the potential of Angiosperms353 for population genomics. bioRxiv:2020.2010.2011.335174.

Smith S.A., Brown J.W. 2018. Constructing a broadly inclusive seed plant phylogeny. Amer. J. Bot. 105:302-314.

Soltis D.E., Smith S.A., Cellinese N., Wurdack K.J., Tank D.C., Brockington S.F., Refulio-Rodriguez N.F., Walker J.B., Moore M.J., Carlsward B.S., Bell C.D., Latvis M., Crawley S., Black C., Diouf D., Xi Z., Rushworth C.A., Gitzendanner M.A., Sytsma K.J., Qiu Y.-L., Hilu K.W., Davis C.C., Sanderson M.J., Beaman R.S., Olmstead R.G., Judd W.S., Donoghue M.J., Soltis P.S. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Amer. J. Bot. 98:704-730.

Soltis D.E., Soltis P.S., Chase M.W., Mort M.E., Albach D.C., Zanis M., Savolainen V., Hahn W.J., Hoot S.B., Fay M.F., Axtell M., Swensen S.M., Prince L.M., Kress W.J., Nixon K.C., Farris J.S. 2008. Angiosperm phylogeny inferred from 18s rDNA, *rbcL*, and *atpB* sequences. Bot. J. Linn. Soc. 133:381-461.

Soltis P.S., Folk R.A., Soltis D.E. 2019. Darwin review: Angiosperm phylogeny and evolutionary radiations. Proc. R. Soc. Lond. B Biol. Sci. 286:20190099.

Soto Gomez M., Pokorny L., Kantar M.B., Forest F., Leitch I.J., Gravendeel B., Wilkin P., Graham S.W., Viruel J. 2019. A customized nuclear target enrichment approach for developing a phylogenomic baseline for *Dioscorea* yams (Dioscoreaceae). Appl. Plant. Sci. 7:e11254.

Toronto International Data Release Workshop Authors. 2009. Prepublication data sharing. Nature 461:168-170.

Van Andel T., Veltman M.A., Bertin A., Maat H., Polime T., Hille Ris Lambers D., Tjoe Awie J., De Boer H., Manzanilla V. 2019. Hidden rice diversity in the Guianas. Front. Plant Sci. 10:1161.

Villaverde T., Pokorny L., Olsson S., Rincón-Barrado M., Johnson M.G., Gardner E.M., Wickett N.J., Molero J., Riina R., Sanmartín I. 2018. Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. New Phytol. 220:636-650.

WCVP. 2020. World Checklist of Vascular Plants, version 2.0. Facilitated by the Royal Botanic Gardens, kew. Published on the internet; http://wcvp.science.kew.org/, retrieved 18 November 2020.

Weitemier K., Straub S.C.K., Cronn R.C., Fishbein M., Schmickl R., McDonnell A., Liston A. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. Appl. Plant. Sci. 2:1400042.

Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E., Matasci N., Ayyampalayam S., Barker M.S., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R., Wafula E., Der J.P., Graham S.W., Mathews S., Melkonian M., Soltis D.E., Soltis P.S., Miles N.W., Rothfels C.J., Pokorny L., Shaw A.J., DeGironimo L., Stevenson D.W., Surek B., Villarreal J.C., Roure B., Philippe H., dePamphilis C.W., Chen T., Deyholos M.K., Baucom R.S., Kutchan T.M., Augustin M.M., Wang J., Zhang Y., Tian Z., Yan Z., Wu X., Sun X., Wong G.K.-S., Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. U.S.A. 111:E4859.

Yan Z., Du P., Hahn M.W., Nakhleh L. 2020. Species tree inference under the multispecies coalescent on data with paralogs is accurate. bioRxiv:498378.

Yang L., Su D., Chang X., Foster C.S.P., Sun L., Huang C.-H., Zhou X., Zeng L., Ma H., Zhong B. 2020. Phylogenomic insights into deep phylogeny of angiosperms based on broad nuclear gene sampling. Plant Commun. 1:100027.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19:153.

1202 Zhao T., Xue J., Kao S.-m., Li Z., Zwaenepoel A., Schranz M.E., Van de Peer Y. 2020.

1203 Novel phylogeny of angiosperms inferred from whole-genome microsynteny analysis.

1204 bioRxiv:2020.2001.2015.908376.

1205

1206

TABLES

Table 1. Total number of angiosperm samples included at three stages of data release preparation. The first column represents all samples available in the initial dataset. The second column indicates samples included in our preliminary tree, prior to family identification validation, but after removal of samples for which the sum of the gene lengths fell below 20% of the median value across all samples. The third column provides numbers for the samples made public in the Kew Tree of Life Explorer, Data Release 1.0, and included in our final phylogenetic tree. Numbers of angiosperm families, genera and species in each data subset are provided in brackets (as families/genera/species).

Data source	Initial dataset	Preliminary tree pre-validation	Final tree and Data Release 1.0
Target sequence capture data	2,522 (304/1988/2397)	2,438 (297/1947/2340)	2,374 (292/1903/2280)
1KP transcriptomes	689 (254/544/682)	678 (250/530/677)	664 (245/517/663)
Annotated genomes	61 (23/43/59)	61 (23/43/59)	61 (23/43/59)
Total	**3,272 (413/2428/3079)**	**3,177 (410/2388/3028)**	**3,099 (404/2333/2956)**

Table 2. Results of validation of sample family identification. The family identification of each sample was scored as confirmed, inconclusive or rejected according to both DNA barcode and phylogenetic validations. Where only a single-family representative was included, samples were tested at the ordinal level. Based on these results, samples were automatically included, excluded, or held for review. See Materials and Methods and Fig. 4 for more details.

		DNA barcode validation		
		Confirmed	**Inconclusive**	**Rejected**
Phylogenetic validation	**Confirmed**	2,666 Include	398 Include	4 Review
	Inconclusive	27[a] Review	7 Review	3 Exclude
	Rejected	8 Review	42 Exclude	22 Exclude

[a]Samples with confirmed family (barcode), but for which the placement cannot be confidently assessed were reviewed.

1232 **Table 3.** Target sequence capture and gene recovery statistics by sample or gene for Data Release 1.0, including the results of mining of genes
1233 from the 1KP and annotated genome datasets. The upper five rows apply to target sequence capture data only.

1234

	Median	Mean	Standard deviation	Minimum	Maximum
Raw reads per sample	1,756,586	2,821,720	3,075,500	16,756	40,535,096
Trimmed reads per sample	1,585,152	2,549,298	2,790,691	13,911	36,051,667
Percentage of reads on-target per sample (across all recovered genes)	5.676	8.020	7.704	0.005	50.953
Read depth per sample (at bases with ≥4x depth across all recovered genes)[a]	38	90	105	5	2,243
Read depth per gene (at bases with ≥4x depth across all samples)[a]	38	97	37	27	226

Recovered genes per sample:					
Target sequence capture data	338	330	24	148	353
1KP transcriptomes	341	328	44	30	353
Annotated genomes	346	341	13	287	353
Recovered genes lengths across all samples[b] (bp):					
Target sequence capture data	387	477	347	48	3,564
1KP transcriptomes	717	803	466	50	4,689
Annotated genomes	972	1,136	642	45	8,601
Sum of recovered gene lengths per sample (bp):					
Target sequence capture data	161,312	157,560	43,545	34,326	256,944
1KP transcriptomes	275,372	262,715	66,593	6,498	367,419
Annotated genomes	390,123	387,630	18,680	321,666	427,322
Percentage length per recovered gene[c] across all samples:					
Target sequence capture data	63	62	16	27	96

1KP transcriptomes	88	85	10	44	100
Percentage length of recovered genes[c] per sample:					
Target sequence capture data	63	62	14	20	95
1KP transcriptomes	88	84	13	16	100

[a]calculated by Samtools depth program

[b]see Supplementary Figure S5

[c]percentage length calculated against each representative target gene

1239 **Table 4**. Properties of the 347 gene alignments and gene trees underpinning the species tree

1240 included in the Kew Tree of Life Explorer Data Release 1.0.

	Median	Mean	Standard deviation	Minimum	Maximum
Number of samples	2,421	2,377.2	358.8	491	3,014
% of total samples[a]	77.9	76.5	11.5	15.8	96.9
Alignment length	1,259.0	1,533.9	985.7	250	8,119
% missing data[b]	58.9	57.9	11.3	14.4	85.8
Variable sites	1,224	1,469.7	940.6	240	7,873
% variable sites	96.6	96.0	2.5	81.5	100
Parsimony informative sites	1,137	1,369.4	859.3	233	6,792
% parsimony informative sites	90.7	90.0	4.20	69.1	98.9
% nodes in gene trees above 30% UFBS[c]	98.9	98.5	1.3	90.7	99.9
Mean support[c] of all nodes	88.1	87.8	2.7	78.9	94.3
Median support[c] of all nodes	98.0	97.6	1.8	90.0	100

1241 [a]percentage of samples in species tree present in alignment/gene tree

1242 [b]percentage of empty cells in each alignment

1243 [c]UFBS: ultrafast bootstrap

1244

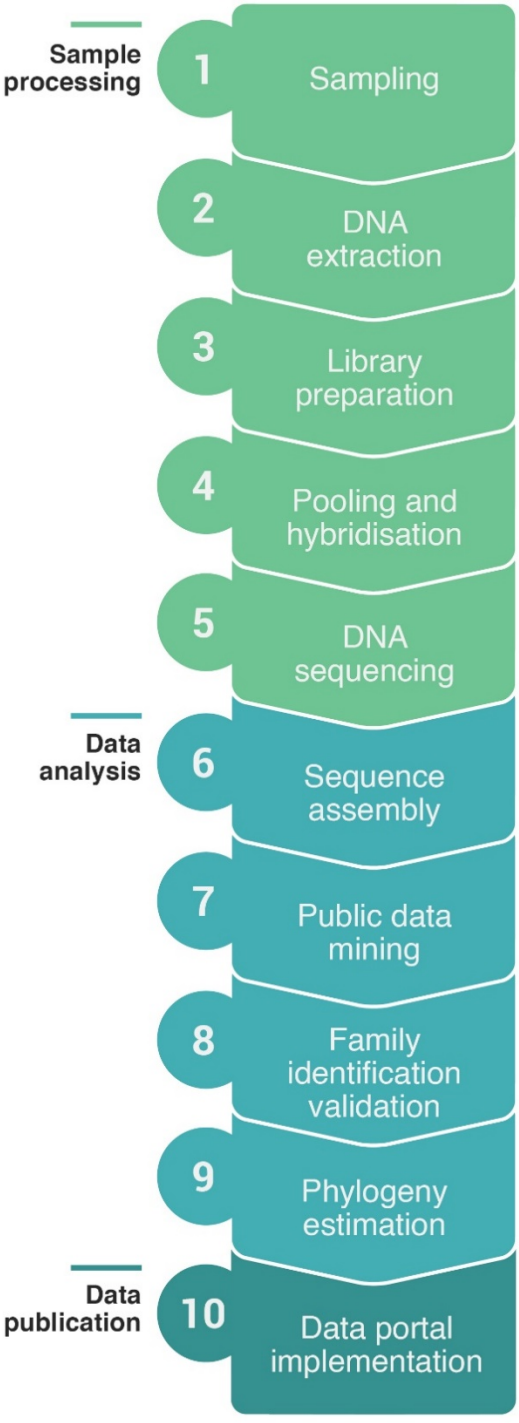
1245

1246 **FIGURES**

1247

1248 **Figure 1**. Summary workflow. Overview of steps taken by the PAFTOL project to generate

1249 Data Release 1.0 of the Kew Tree of Life Explorer (https://treeoflife.kew.org).



1250

Figure 2. Sample processing workflow. Processes are indicated by bold headings with reagents and machines used given below. Quality control checkpoints are indicated in dark grey boxes.

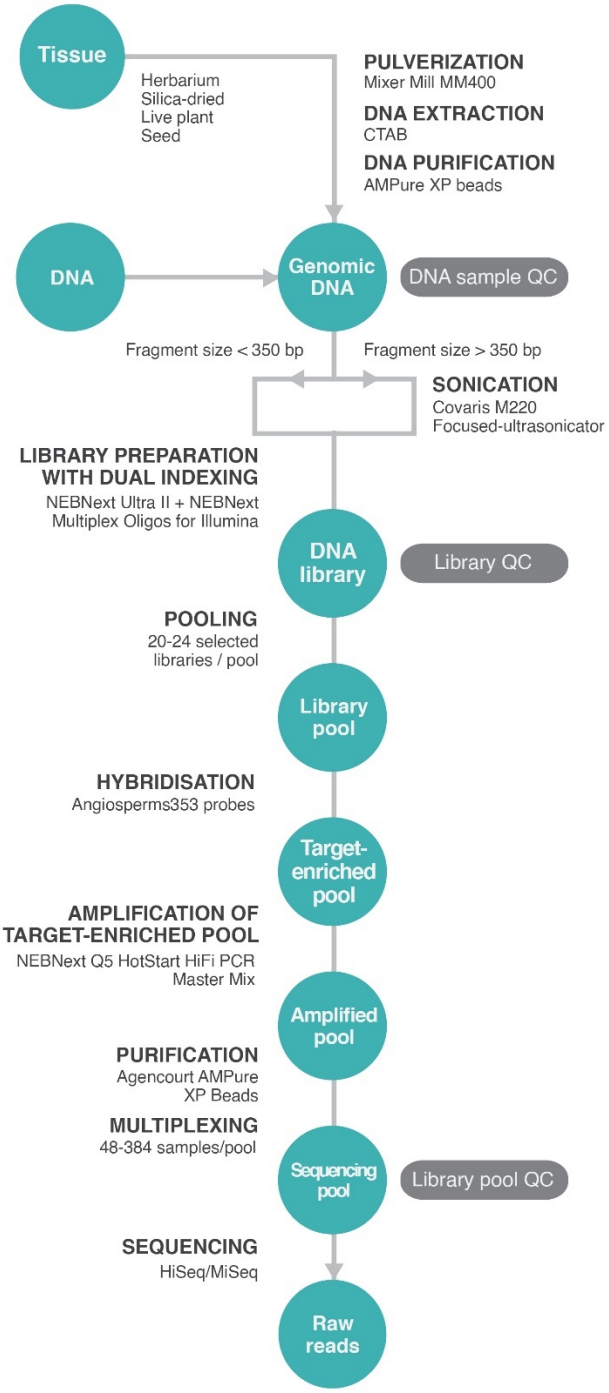


Figure 3. Data analysis workflow. Pipeline products are shown in blue-green circles (available to download via the Kew Tree of Life Explorer, https://treeoflife.kew.org). Processes are indicated by bold headings with programs used given below.

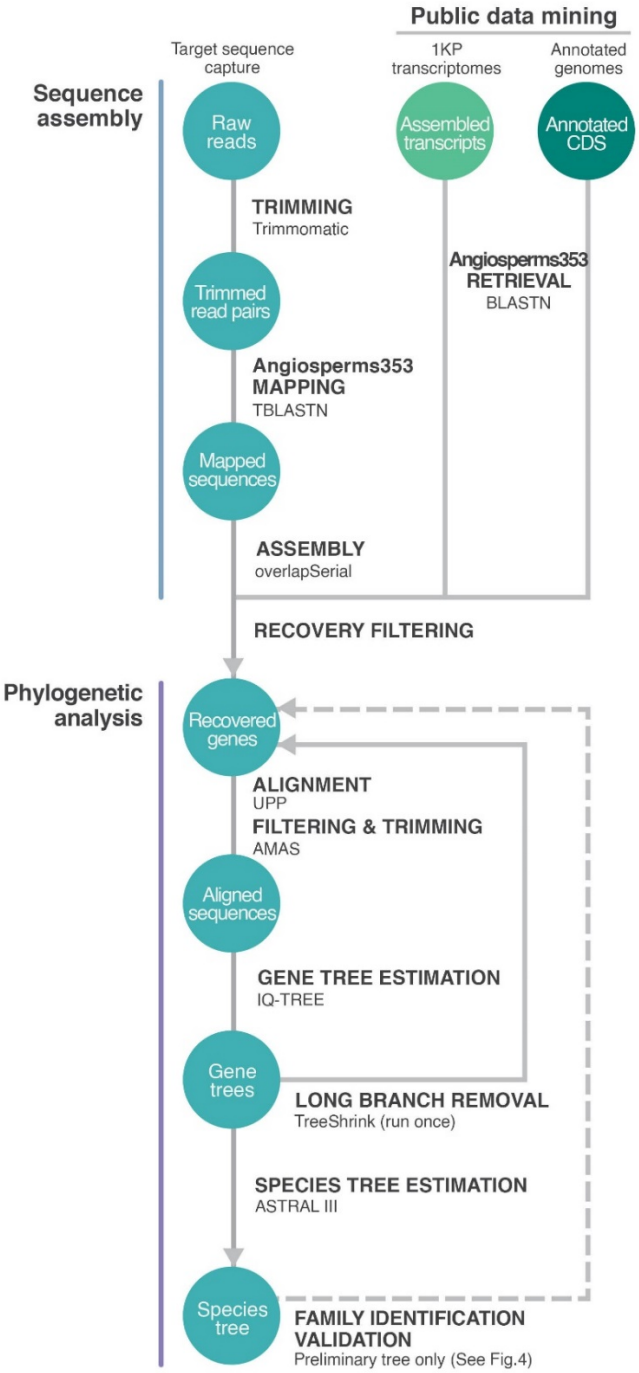


Figure 4. Family identification validation workflow. Processes are indicated by bold headings. Embedded table (bottom right) indicates decisions made for each sample based on the two validation steps.

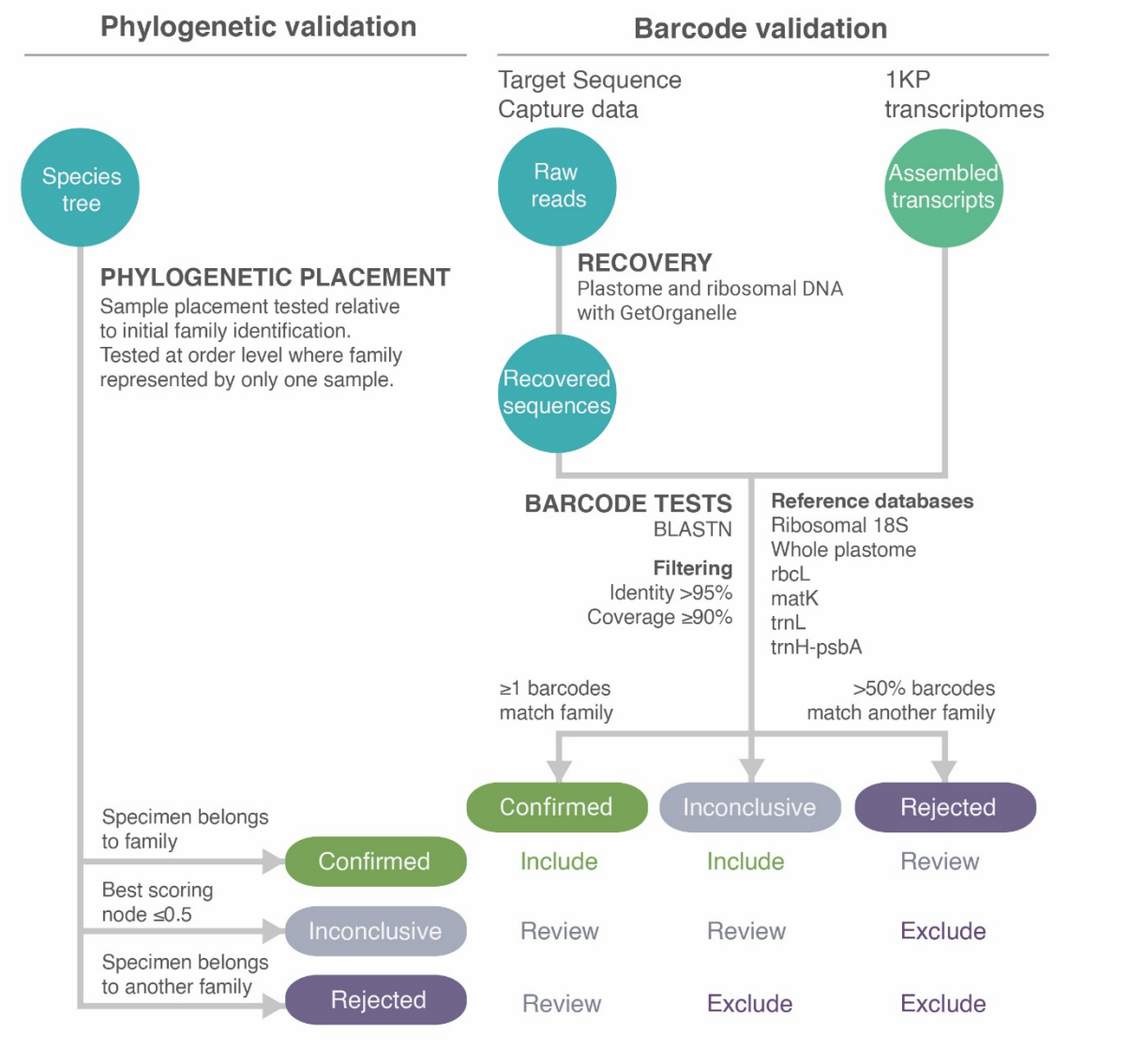


Figure 5. Data publication workflow. Implementation of Kew Tree of Life Explorer data portal is illustrated. Arrows indicate data flow from internal repository to public interface. Infrastructural components are shown in purple; publicly available information is shown in green. External links available from the portal are listed in the lower left.

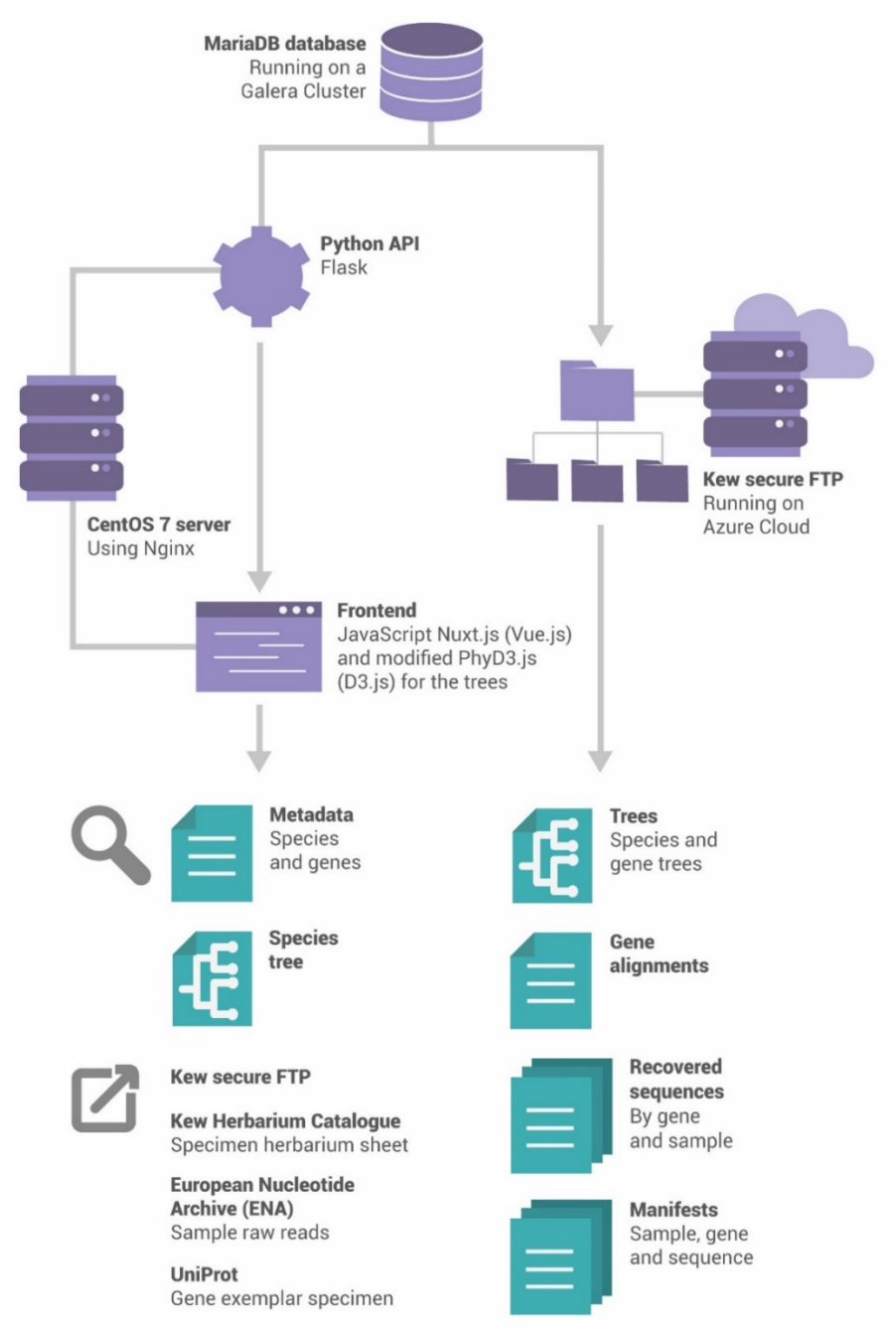
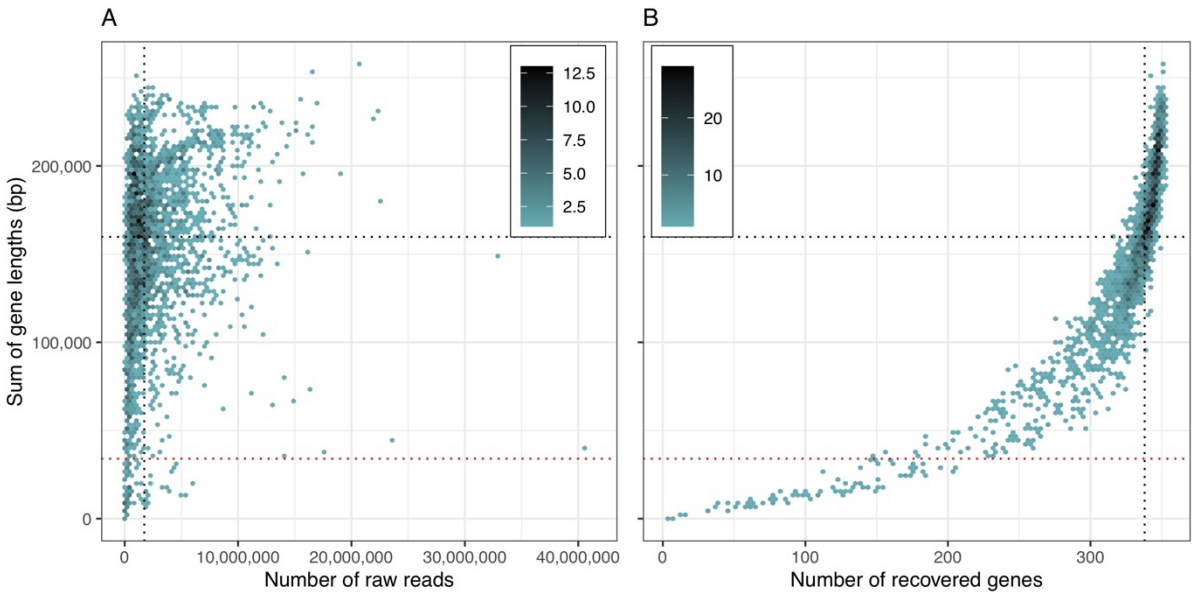
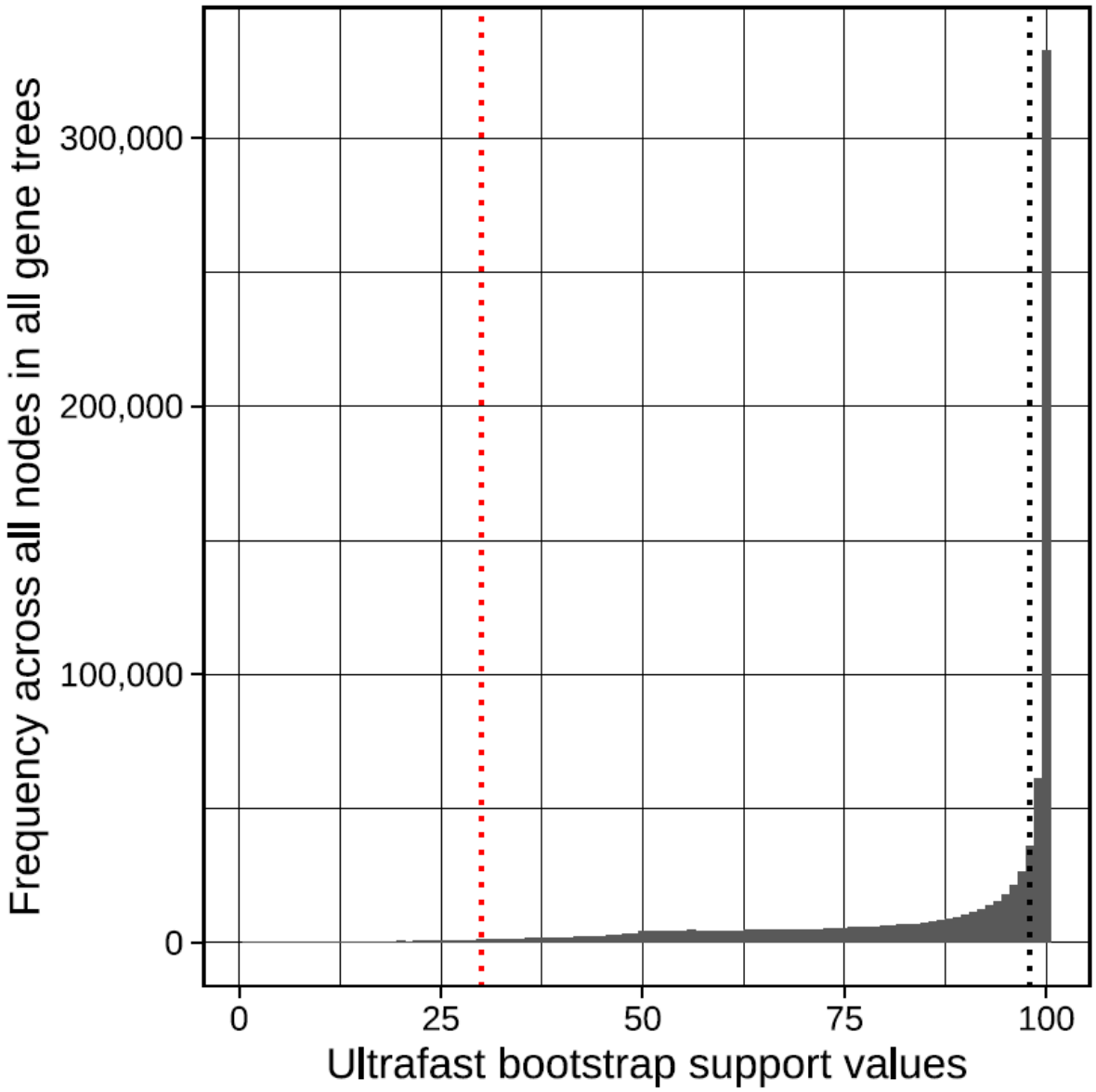


Figure 6. Density plots of target sequence recovery from our raw data. Data are presented prior to any filtering, illustrating relationships of sum of gene lengths (bp) to (a) the number of raw reads and (b) the number of recovered genes. Colours indicate density of data points. Black dotted lines indicate medians of variables and red dotted lines indicate the threshold used to remove samples from downstream analyses, set as 20% of the median value across all samples.



1282 **Figure 7**. Distribution of ultrafast bootstrap support values across all nodes in all gene trees.
1283 Bootstrap values were estimated with IQ-TREE 2.0.5 (Hoang et al. 2017; Minh et al. 2020).
1284 Black dotted line indicates the median (98%) and the red dotted line indicates the threshold
1285 (30%) for collapsing nodes with low support prior to species tree inference with ASTRAL-III
1286 (Zhang et al. 2018). Only 1.3% of all nodes across gene trees are collapsed prior to species
1287 tree inference.
1288



1289
1290

Figure 8. Summary of node properties in the species tree derived from ASTRAL-III (Zhang et al. 2018). Data are grouped by (a, c, e) taxonomic level and (b, d, f) major taxonomic groups. In a, c and e, "within families" refers to relationships within families; "among families" refers to relationships within orders but among families; "among orders" refers to relationships among orders. Box plots show medians, 1st and 3rd quartiles (hinges), and the full distribution excluding outliers (whiskers).