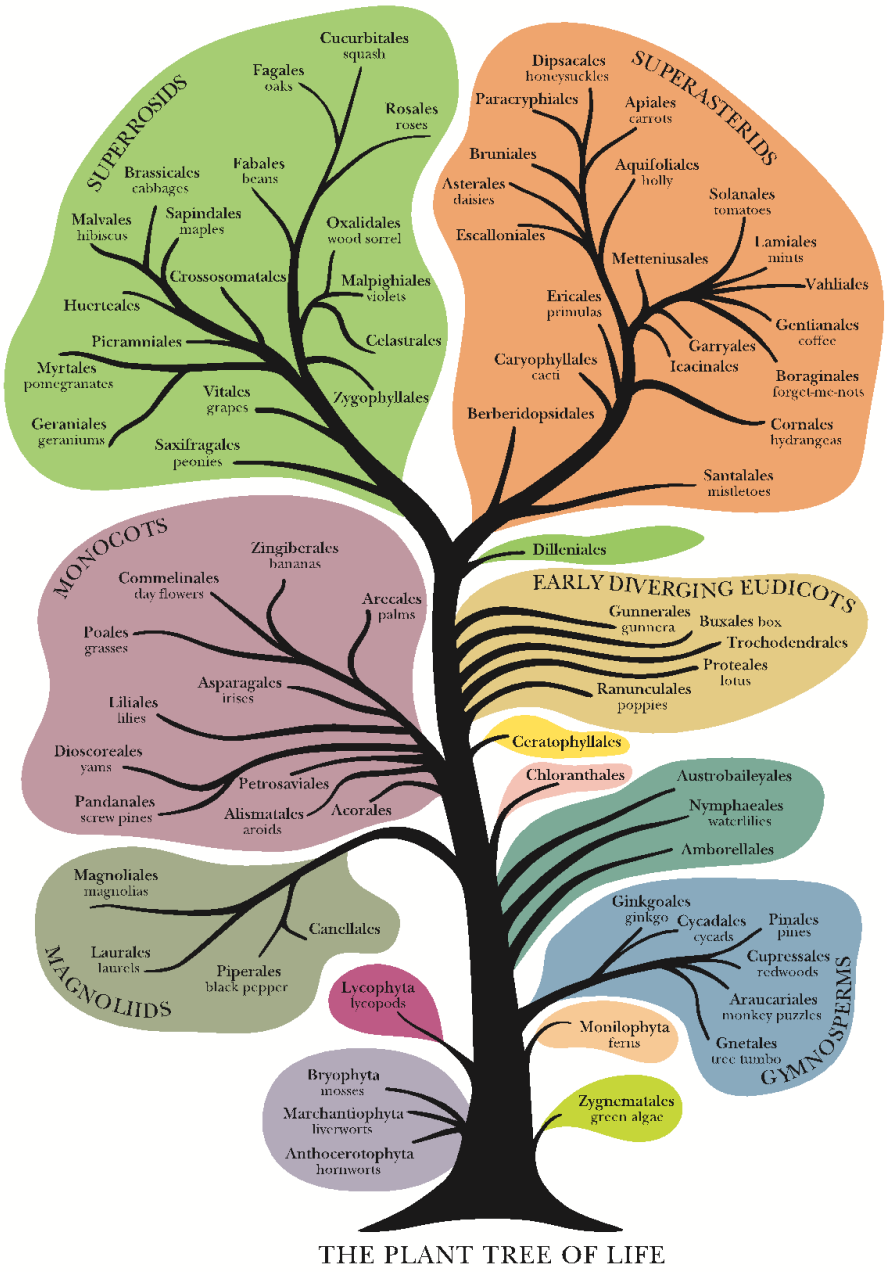
PAFTOL – Plant and Fungal Trees of Life

Third Annual Report



THE PLANT TREE OF LIFE

Prepared by: William Baker, Vanessa Barber, Felix Forest, Ilia Leitch, Jan Kim, Olivier Maurin and Wolf Eiserhardt.

Project Summary

Evolutionary trees are powerful tools for prediction, species discovery, monitoring and conservation. To better understand how the world's plants and fungi are related to each other and how they have evolved, we aim to complete the Plant and Fungal Trees of Life (PAFTOL). Through comparative analysis of DNA sequence data, the backbones of these Trees of Life are already relatively well understood, and many components have been studied in detail. However, DNA data are still lacking for many genera and the vast majority of species of plants and fungi, preventing their accurate placement within this evolutionary framework.

To complete the Plant and Fungal Trees of Life for all genera, we will utilise Kew's collections to produce genome-scale DNA data for a representative of each genus of plant and fungi using high-throughput sequencing technologies. This comprehensive investigation of evolutionary relationships will provide a unifying framework for comparative plant and fungal research, greatly accelerating the discovery of new taxa, particularly in less well-known groups, as well as facilitating the exploration of properties and uses. The project is an essential step towards the compilation of genomic data for all known species using Kew's collections.

Project Objectives

Kew is committed to completing the Plant and Fungal Trees of Life by generating and compiling high throughput sequencing data for one representative of all 14,000 flowering plant genera and all 8,200 fungal genera by 2020.

Executive summary

The completion of the Plant and Fungal Trees of Life is a strategic science priority for the Royal Botanic Gardens, Kew (RBG Kew). The Plant and Fungal Trees of Life Project (PAFTOL) has received significant funding from the Calleva Foundation, the Sackler Trust and the Garfield Weston Foundation. This funding has allowed RBG Kew to make significant headway into achieving PAFTOL's objectives.

In its first two years, April 2016 – March 2018, the project team focused on the plant component of PAFTOL, and in putting its operational foundations and infrastructure in place. Most notably this included:

- Establishing the Sackler Phylogenomics Laboratory and the Calleva Phylogenomics Research Programme.
- Establishing methods, protocols, and systems for processing the samples through the two work packages.
- Engaging the broader Kew community through training and support for 30 subprojects.
- The design and implementation of a novel genomic protocol for the delivery of PAFTOL Plants, which uses molecular probes ("PAFTOL baits") to "fish out" 353 genes ("PAFTOL genes") for studying evolutionary relationships.
- The completion of a pilot study that trialled the PAFTOL baits protocol, and a pilot study for fungal genera.
- Development of bioinformatics software to process PAFTOL genomic data.
- The publication of three scientific papers, including one on the PAFTOL baits development.
- Building strong links with key external stakeholders, especially in the USA. And the presentation of PAFTOL at major international conferences, workshops and seminars.
- Through PAFTOL's leadership, the PhyloSynth global network of researchers for the synthesis of plant tree of life data was established.
- PAFTOL pursued opportunities at Kew to engage visitors to the gardens with the science of the project, including a PAFTOL activity stand at the Kew Science Festival.

Activities in reporting year 2018-2019

- The project completed a 20% coverage of flowering plant genera by March 2019, equivalent to 2,800 sequenced specimens. Exceeding the target number of genera laid out in the funding agreement by 40%.
- The PAFTOL bait kit 'Angiosperms-353' was made commercially available through Arbor Bioscience and a publication on the bait sequences and the methods used to develop them was also published in the high impact journal *Systematic Biology*.
- A review of lab processes and techniques was completed, including an assessment of current technology and time-limiting steps, and ways to reduce the cost per sample.
- PAFTOL secured a major contract with a genomic sequencing company in order to scale up components of the lab work and reduce costs.
- The team increased its bioinformatics capability, with input from four newly appointed staff - The Head of Genomics at Kew, Dr. Paul Kersey, two post-doctoral researchers, and a bioinformatician.
- The PAFTOL gene recovery software 'PAFtools' was developed and enhanced.
- Work began on the plastid recovery pipeline which will complement the data retrieved from the 353 nuclear genes.
- The Data Analysis work package supported the wider PAFTOL research community with analyses of sequence data which will result in at least three publications in 2019-2020.
- The Agaricinae sub-project of the fungal pilot was completed.
- PAFTOL pursued and developed collaborative opportunities at both a national and international level, including a number with key external collaborators who are leaders in the fields of systematics, phylogenomics and bioinformatics.
- Staff participated in 11 high-profile conferences, seminars and workshops, and presented the project to a wide, international audience.
- PAFTOL hosted its first annual symposium to celebrate and share the results and early research findings being generated by the project.
- PAFTOL delivered talks and activities to the public at Kew Gardens and Wakehurst, and pursued opportunities to enhance the project's visibility across the two sites including through the interpretation being developed for Kew's Evolution Garden - a major new garden for 2019.

In Year 4, PAFTOL will broaden its ambitions and aim for a 50% coverage of flowering plant genera by the end of the project, May 2021, equivalent to 7,000 sampled specimens. This year PAFTOL will capitalise on its increased phylogenomic and bioinformatic capabilities and processing capacity to drive forward novel research and meet the projects ambitious targets. This step-change will firmly cement the reputation of PAFTOL as a trailblazer in the field of botanical phylogenomic research and enable it to meet its pioneering aims. PAFTOL will pursue and develop collaborative opportunities at both a national and international level and promote its activities to both the scientific community and the public, ensuring that the impact of the project continues to grow.

PAFTOL Team

William Baker
Head, Comparative Plant & Fungal Biology,
PAFTOL Project Leader and PI

Felix Forest
Senior Research Leader,
PAFTOL PI

Paul Kersey
Deputy Director of Science (Bioinformatics and Genomics),
PAFTOL PI

Ilia Leitch
Assistant Head, Comparative Plant & Fungal Biology,
PAFTOL PI

Wolf Eiserhardt
Senior Research Leader to (8/17), Honorary Research Associate

Olivier Maurin
Senior Research Leader,
Work Package 1 Leader

Grace Brewer
Research Assistant,
Work Package 1

Niroshini Epitawalage
Research Assistant,
Work Package 1

Robyn Cowan
Lab technician

Jan Kim
Senior Bioinformatician,
Work Package 2 Leader

Paul Bailey
Bioinformatician,
Work Package 2

Noor Al-Wattar
Bioinformatics Intern,
Work Package 2

Jim Clarkson
Phylogenomics
Post-Doctoral Researcher

Alexandre Zuntini
Phylogenomics
Post-Doctoral Researcher

Vanessa Barber
Project Manager

Lisa Pokomy
Garfield Weston Phylogenomics Research Fellow

Recruitment and volunteers

In Year 3, between September and October 2018, three recruitments were made; two Postdoctoral Researchers, Dr. Jim Clarkson and Dr. Alexandre Zuntini, and Bioinformatician Dr. Paul Bailey.

The project has also attracted two highly skilled volunteers, Dr. Catherine McGinnie and Dr. Haydn Thompson who assist with laboratory duties related to the Data Production work package. In addition, the project has gained a sandwich intern student from the University of Plymouth, Noor Al-Wattar, who is supporting Dr. Jan Kim with the Data Analysis work package.

Data Production work package

The remit of the Data Production work package is to source and process specimens in order to produce genomic quality DNA sequences of all flowering plant genera. The sequence data produced by Data Production feeds into the Data Analysis work package for synthesis, analysis and dissemination. Data Production is led by senior researcher Dr. Olivier Maurin, and supported by research assistants, Dr. Niroshini Epitawalage and Grace Brewer, who process DNA samples and prepare DNA libraries in the laboratory ready for sequencing.

In the first two years Data Production established the foundations required to deliver this work package efficiently:

- Created a detailed plan for securing DNA samples of the world's ca. 14,000 flowering plant genera.
- Developed the standards and protocols for sampling living plant material to ensure consistent quality and optimal preservation.
- Developed a lab methodology and procured the required equipment to produce genomic data from DNA samples. Including the design and testing of the PAFTOL bait kit.
- Sourced, generated and sequenced genomic data for all 416 flowering plant families for the PAFTOL pilot project.
- Harnessed expertise and engagement across Kew through training and support for 30 subprojects, each of which will result in at least one scientific paper.

Progress in Year three

1. **Processed samples and produced sequence data to meet the 25% target of 3,500 samples by March 2019.**

   In year three the project entered 'production mode', aiming to generate sequence data for 25% of the 14,000 angiosperm genera (3,500 sequenced samples), by March 2019. The team have generated 2,800 to date, with a further 400 samples prepped for sequencing in the lab work flow to follow soon. This number, although short of the 25% target, remains a significant achievement – exceeding the number of sequences agreed under the terms of the funding, 2000, by 40%.

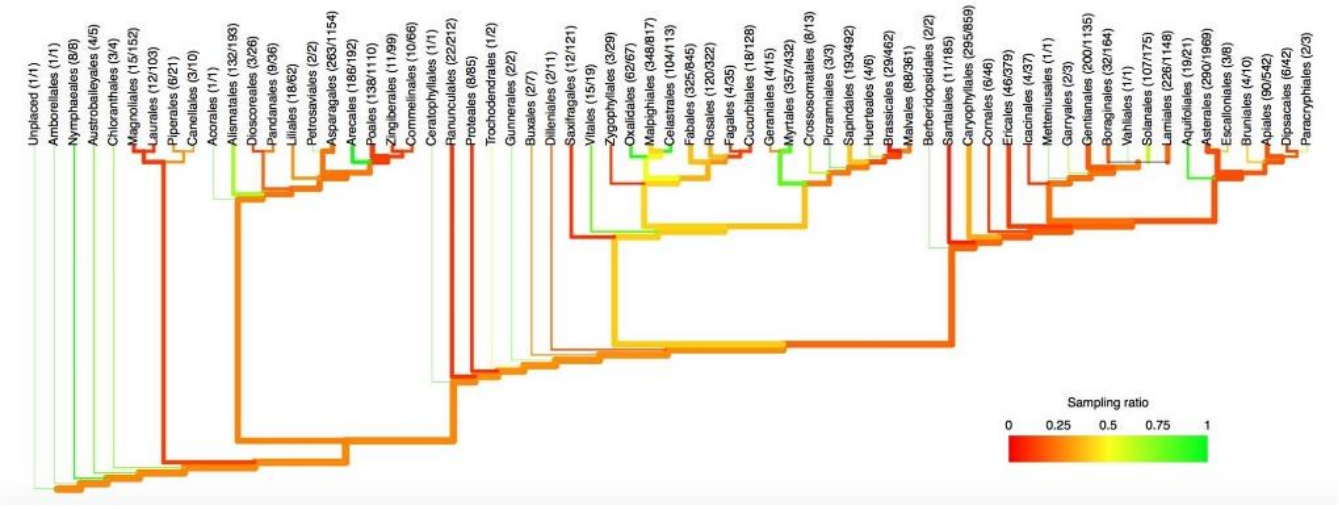


Figure 1: Evolutionary tree of relationships between different groups in the Angiosperms. The tree reflects PAFTOL's sampling across these groups to date. The branch width is proportional to the total number of genera in that group, and the colour represents the sampling ratio, i.e: the % of that group which have been sampled.

2. **The PAFTOL 'Angiosperms- 353' bait kit.**

Baits are short molecular (RNA) probes which bind to specific target genes and can be used to "fish out" those targets. The PAFTOL baits can fish out a maximum of 353 genes, and it is through a comparative analysis of these genes that the plant trees of life can then be constructed. The PAFTOL bait kit is unique because it can be used in any flowering plant (angiosperm) family, in contrast to other kits which are typically very specific to a single family. In year three PAFTOL published the bait kit and the methodology behind its design in the high impact journal *Systematic Biology*. Link to the paper: https://doi.org/10.1093/sysbio/syy086 It also made the bait sequences and kit commercially available to the scientific community through Arbor Biosciences and published a paper on the baits and methods used to develop them. This will greatly increase the worldwide impact of the PAFTOL's work.

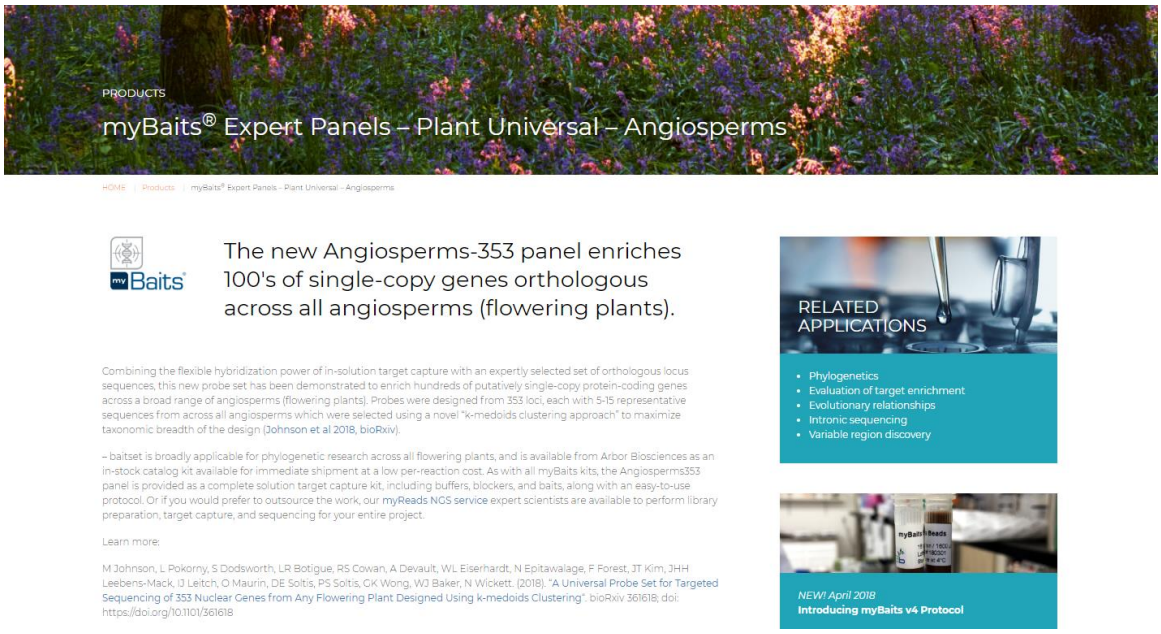


Image 1: A screenshot of the Arbor Bioscience webpage advertising the Angiosperms-353 bait kit.

3. **Reviewed lab processes and techniques, including an assessment of current technology and time-limiting steps, as well as factors affecting sample success.**

This year, with the ambitious 25% target in mind, the Data Production Team invested significant time in reviewing the methods used for processing samples for time and cost savings, balanced against an understanding of the factors affecting quality. This review identified several steps where reducing the volume of certain expensive chemicals and reagents could still result in good quality DNA sequences being produced. This not only enabled the team to reduce its per sample costs by one third, but also to build an informed projection of the number of samples the project can deliver through to the end of the funding period. This review process will be part of an on-going cycle of improvement and there is further work to do on this in year four.

4. **Established an outsourcing contract to scale up the production of sequence data.**

In year three as the project transitioned from the pilot stage through to 'production mode' it was imperative to secure a sequencing contract to scale up the generation of sequence data in order make this process both fast and cost effective. The project has now secured a two-year contract with a sequencing supplier, Macrogen, which will use its NovaSeq6000 sequencing platform. The benefit of this platform is that it generates an order of magnitude more read data than the MiSeq platform available in-house at RBG Kew, and at a quarter of the cost per sample. Preparation of samples for sequencing is still conducted in-house at Kew.

Data Analysis work package

The completion of PAFTOL rests upon the availability of bioinformatic pipelines and computing infrastructure to process genomic sequence data reliably, at scale, with minimum human input. The purpose of Data Analysis, led by senior bioinformatician Dr Jan Kim, and supported by Dr. Paul Bailey and intern Noor Al-Wattar, is to process the sequencing data generated by Data Production into a comprehensive tree of life for plants.

In the first two years Data Analysis established the infrastructure and tools required to deliver this work package efficiently, including:

- The procurement of a computer cluster needed to analyse PAFTOL's genomic data.
- The development of a software framework and toolkit 'PAFtools' to analyse the genomic sequence data produced by the Data Production work package.

1. Angiosperm-353 sequence recovery: PAFtools software developed and enhanced.

   In year three the Data Analysis team spent time optimising the PAFtools software that recovers the 353 PAFTOL genes from the DNA sequence data. The initial design was inspired by the "HybPiper" pipeline, which was developed by colleagues at the Chicago Botanic Gardens. However, in year three areas of the code were extensively rewritten to automate all 'housekeeping' aspects, such as creating temporary directories and removing them when finished, as well as making significant improvements to the pipeline's ability to detect gene sequences, especially where these are highly divergent from the known reference sequences.

2. Development of a plastid recovery pipeline.

   PAFTOL currently constructs its tree phylogenies based on nuclear gene sequences (DNA recovered from the nucleus of the cell). However, it is also possible to retrieve and integrate plastid data, that is, genetic information retrieved from other cell organelles such as the chloroplast. Plastid DNA is interesting because the way that it is inherited can cast a different interpretation on plant evolutionary relationships. Combining both plastid and nuclear genes will allow us to be more confident in the results and relationships revealed by the phylogenetic trees. In addition the recovery of specific plastid genes such as rbcL and matK, which have been used extensively over the last 10-15 years in phylogenomics, will allow PAFTOL to connect back to old datasets enhancing the value of the data produced by the project As a result in year three, PAFTOL began work on a plastid recovery pipeline which will be integrated into the nuclear analysis pipeline.

3. Supported the PAFTOL research community with analyses of sequence data.
   The support provided by the Data Analysis work package to both core PAFTOL staff and other collaborators is hugely important for the generation of high quality, repeatable and informative phylogenetic trees, and the research questions which they help to address. In year three the Data Analysis team helped to support a great number of sub-projects, which will result in several publications over the coming year. Initial results from some of these studies are expanded on in the research results section of this report.

Research results from 2018-2019

In year three the project completed half of its 30 'Expression of Interest' sub-projects, collaborative projects between PAFTOL and the wider scientific community on specific clades, orders or families of plants that align with PAFTOL's big picture goal to sample all 14,000 angiosperm genera. Here are some early results from two of these sub-projects. Each of which will result in at least one publication.

The Fungal Pilot

Although the PAFTOL project has thus far primarily focused on constructing the plant tree of life, progress has been made towards completing a fungal pilot too. Fungi are a particularly fascinating group as the evolutionary relationships between them are far more poorly understood than those of plants.

Agaricales is the largest order of mushroom-forming Fungi. The suborder Agaricineae, contains some of the most familiar types of mushrooms, from the ubiquitous common mushroom to the hallucinogenic fly agaric and the bioluminescent jack-o-lantern mushroom.

In year three the fungal pilot team focused on the Agaricinae sub project. They generated 24 fungal genome sequences from the group and combined these with 11 publicly available genome sequences to create a phylogenetic tree. The results from this study will help to resolve the evolutionary relationship between genera in this family.

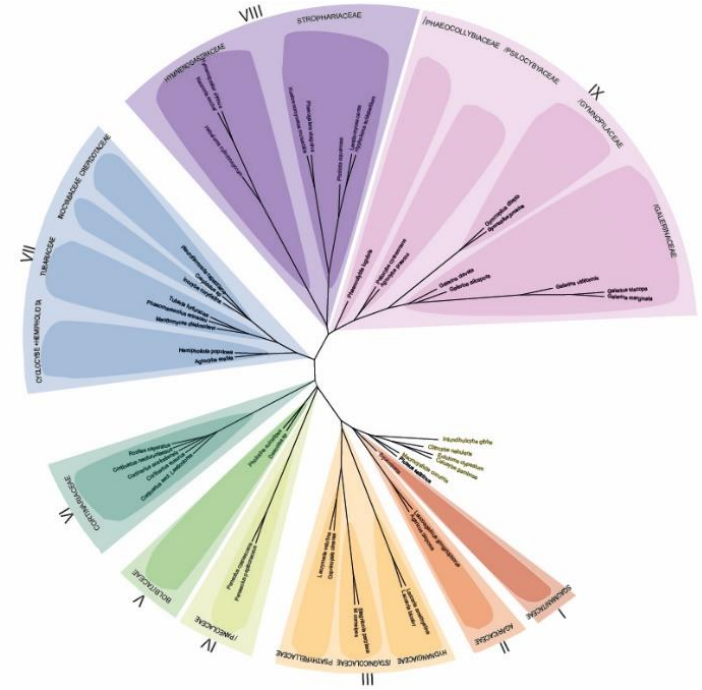


Figure 2: A phylogenetic tree created with ASTRAL-II of the Agaricineae.

Resolving evolutionary relationships between genera in Cyperaceae

The Cyperaceae (or sedge family) occupy a broad range of habitats, from Arctic tundra to tropical forests. They are perceived as rather uniform and grass-like, but sedges are phenotypically and ecologically very diverse, and include aquatic plants, ephemerals, dwarf shrubs, and tall perennials with large and complex flowers. Cyperaceae have a worldwide economic significance with about 10% of species having a commercial application. But, despite this, systematic relationships in Cyperaceae at the level of the genera are not fully resolved. The Cyperaceae project compared a well-regarded phylogenetic tree for the family based on Sanger sequencing data, with a tree created using Angiosperm-353 data from 114 samples to inform a new DNA-based classification. A comparison of the two trees highlighted areas of agreement and conflict in the placement of certain sub-families, and confirmed the hypothesized placement of a species called Koyamaea neblinensis into the sub-family Cyperoideae,

Image 2: Illustration of *Koyamaea neblinensis*. Based on morphological data, *Koyamaea* was placed it in its own tribe, Koyamaeae, but this was not confirmed until this study.

Stakeholder and public engagement

This year PAFTOL strengthened its relationships with key individuals in the phylogenomic world, such as Doug and Pam Soltis (University of Florida), Jim Leebens-Mack (University of Georgia), and Norm Wickett and Matt Johnson (Chicago Botanic Garden) through co-publication of the Angiosperms353 bait kit in *Systematic Biology*.

In 2016 PAFTOL participated in the founding workshop of the Earth BioGenome Project (EBP) at the Smithsonian Institution, an ambitious "moon-shot" vision to sequence the genomes of all species of life on Earth. Subsequently, Kew has participated in the EBP's working group, resulting in a position paper published in the *Proceedings of the National Academy of Sciences of the USA*. Kew has become a key player in this arena on account of its ambition stated through the PAFTOL project and is now widely recognised as a central counterpart in comparative plant genomics in the UK.

The PAFTOL team participated in 11 major conferences presenting the project to a wide, international audience. Most significant among these were, Botany (Minnesota, USA), Evolution (Montpellier, France) and Monocots (Natal, Brazil) at which talks were given on PAFTOL's vision and methods. These talks attracted significant attention and buy-in from other scientists. Many of whom have since sought out collaborative ways of working with PAFTOL, which will help the team to meet its ambitious sampling targets. As a result, some key new stakeholders in the project include the pre-eminent phylogeneticists Stephen Smith (University of Michigan) and Mark Simmons (Colorado State University) and a collaboration with a major programme the Genomics for Australian Plants project.

PAFTOL has been active in sharing its results within the scientific community at Kew. In February 2019 the project hosted its first annual symposium to share the results and early research findings being generated by the project. The symposium was attended by over 120 researchers, students, and visitors and featured 9 short talks from PAFTOL staff and its collaborators.

PAFTOL has pursued many of the rich opportunities available at Kew to engage visitors to the gardens with the science of the project. Namely this has included a PAFTOL activity stand at the Kew Science Festival, an evening presentation to the Friends of Kew, and collaboration with Kew's Learning and Participation Team to deliver original science communication events.

Notably PAFTOL is also working closely with Kew's Interpretation team to develop the messages and stories being told in Kew's new Evolution Garden - a major new garden for 2019 that will profile PAFTOL explicitly. This garden will be laid out according to the contemporary understanding of plant evolutionary relationships. It will tell the story of plant classification through the ages. and how, thanks to step changes in our ability to process and analyse genetic information, botanists, like those on the PAFTOL project, are now able to classify plants according to their DNA sequence, rather than just the physical characteristics of the plant.

Image 3: Kew staff demonstrating how to extract DNA from plants at the PAFTOL stand, Kew Science Festival, 2018.

Plans for 2019-2020

General

- Submit three research papers for publication, including a paper to a major scientific journal for the plant pilot project.
- Further stakeholder engagement with the general public, including a PAFTOL stand at the Kew Science Festival, and though interpretation in the new Evolution Garden.
- Further stakeholder engagement with the scientific community through participation in at least three major international conferences.
- PAFTOL website developed and available to use by scientific community, including development of the PAFTOL Explorer functionality such as data download functionality.

Data Production work package

- Source and process all samples required, and deliver sequence data for, 50% of all flowering plant genera (7,000) by the end of the project.
- Balance the sampling for the 50% target to ensure that the sampling is taxonomically representative.
- Genomic sequence data secured for all 150 species that provide virtually all human food.
- Review high throughput sequencing procedures for efficiencies, including an assessment of current technology and time-limiting steps, as well as factors affecting hybridisation success.

Data Analysis work package

- Version one of the pipeline for genomic data assembly developed and publicly available.
- PAFTOL genomic data made available and accessible to the scientific community to help build the tree of life.