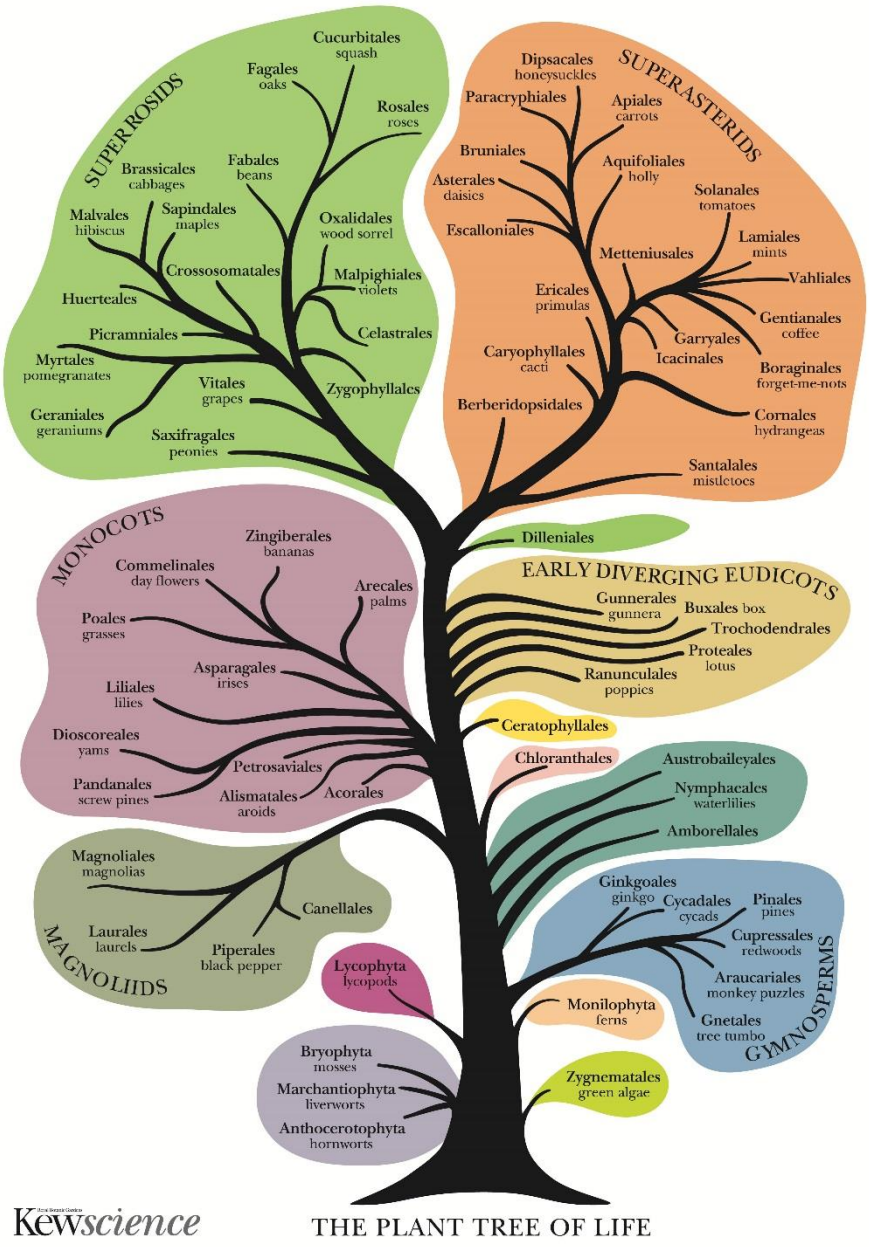
PAFTOL – Plant and Fungal Trees of Life

Royal Botanic Gardens Kew

Second Annual Report



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

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 great 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 allowed RBG Kew to commence work in earnest on the plant component of PAFTOL in April 2016. In year one PAFTOL focused on putting its operational foundations in place. Most notably this included:

- Establishing the project's governance and structure and recruiting project staff.
- Establishing the Sackler Phylogenomics Laboratory and the Calleva Phylogenomics Research Programme, including the procurement of many pieces of specialist equipment and training of PAFTOL staff.
- Establishing baselines, protocols, and systems for processing the samples through the three work packages.
- Engaging the broader Kew community through training and support for 19 subprojects aligned with individual research programmes.
- Building strong links with key external stakeholders, especially in the USA.

The key achievements of PAFTOL during year two (April 2017-March 2018) are as follows:

- Recruitment of two further project staff (a technician and a project manager), hired and in post between August 2017 and October 2017.
- Standards and methods for DNA sample selection from Kew's collections established and trialled and shared with the broader Kew community.

- Ongoing engagement with staff across Kew enabling initiation of a wide range of PAFTOL research on priority plant groups for Kew Science.
- 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 tests the PAFTOL baits protocol on 288 samples to test this ground-breaking approach, resulting in a brand-new plant tree of life in which all flowering plant families are represented.
- Progress on a pilot study of fungal genera, which aims to sequence 100 fungal genomes.
- Development of bioinformatics software to process PAFTOL genomic data.
- Presentation of PAFTOL at major international conferences, including at a successful workshop at Kew in May 2017 and the International Botanical Congress in Shenzhen, China in July 2017.
- Through PAFTOL's leadership, the PhyloSynth global network of researchers for the synthesis of plant tree of life data was established. The inaugural meeting held at Kew involved key global players and resulted in PAFTOL's first scientific paper.

In Year 3, PAFTOL will broaden its ambitions and aim for 25% coverage of flowering plant genera by March 2019, equivalent to 3,500 sampled specimens, exceeding the target number of genera laid out in the funding agreement by ca. 100%. This will put us on course to complete the Plant Tree of Life by generating and compiling sequencing data for one representative of all 14,000 flowering plant genera by 2020, depending on funding. This year PAFTOL will expand its processing capacity further by seeking efficiencies through outsourcing routine components of lab work, while further growing in-house expertise for genomic data generation from Kew's unique collections. In parallel, the team will increase its bioinformatics capability, with input from the newly appointed Head of Genomics at Kew (starting July 2018). These step-changes 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 the widest possible audience, ensuring that the impact of the project continues to grow.

Activities in reporting year 2017-2018

Funding

The PAFTOL programme received commitments for significant funding - £2m from the Calleva Foundation and £1m from the Sackler Trust in 2016. In addition, two postdoctoral positions were established under the Garfield Weston Global Tree Seed Bank Project, which contribute to the objectives of PAFTOL. This funding has allowed Kew to establish the Calleva Phylogenomics Research Programme in the Sackler Phylogenomics Laboratory. So far, the funding has been primarily focused on the delivery of the plant components of PAFTOL, although a fungal pilot project has also been initiated, and both plant and fungal phylogenomic research at Kew are benefiting from the infrastructure and expertise that this funding is facilitating. We continue to seek the further funding required to achieve the ambitions of PAFTOL in full.

PAFTOL team and recruitment:

PAFTOL operates under the following structure:

The PAFTOL Steering Group: This group, chaired by PAFTOL Leader William Baker, is the rudder of the project. It meets monthly to oversee and drive the delivery of PAFTOL, and ensure innovation. It monitors progress against objectives and oversees PAFTOL resourcing and spend. The Steering Group is attended by the Principal Investigators (PIs) of PAFTOL and the PAFTOL Project Manager. The Steering Group reports regularly to the Head of Science, Prof. Kathy Willis, and to the Kew Foundation.

PAFTOL Work Packages: The research of PAFTOL is divided into three discrete but linked work packages (WPs) – WP1 Sampling, WP2 Phylogenomics, and WP3 Bioinformatics. Each WP has a team around it, which meets flexibly as required, to support the delivery of specific objectives within the WP, promoting communication and efficient working. Each WP team is led by the relevant senior project staff member in collaboration with key core staff.

In Year 2, between September and October 2017, two recruitments were made:

- Grace Brewer – Research assistant (WP1&2)
- Vanessa Barber – Project Manager

The project has also attracted two highly skilled volunteers, Alexander Bowles and Dr Beata Klejevskaja who both volunteered part-time for 6 months, assisting with laboratory duties related to WP1 and WP2.

Key Players in PAFTOL

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

Felix Forest
Senior Research Leader
PAFTOL PI

Abigail Barker
Head, Biodiversity Informatics & Spatial Analysis, PAFTOL PI

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

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

Vanessa Barber
PAFTOL Project Manager

Olivier Maurin
Senior researcher
Sampling - WP1 Leader

Steven Dodsworth
Senior researcher
Phylogenomics - WP2 Leader

Jan Kim
Senior bioinformatician
Bioinformatics - WP3 Leader

Grace Brewer
Research assistant (WP1&2)

Niroshini Epitawalage
Research assistant (WP1&2)

Ester Gaya
Research Leader, Mycology

Lisa Pokorny
Garfield Weston
Phylogenomics Research Fellow

Sidonie Bellot
Garfield Weston
Phylogenomics Research Fellow

Robyn Cowan
Lab Technician

Sampling - Work Package 1

The remit of WP1 is to source genomic quality DNA of all flowering plant genera, which is then processed further by WPs 2 and 3. Work Package 1 is led by senior researcher Dr. Olivier Maurin, and supported by research assistants, Niroshini Epitawalage and Grace Brewer, who work across WPs 1 and 2, processing DNA samples and preparing DNA libraries for sequencing.

In year one WP1 established a detailed road map for securing DNA samples of the world's ca. 14,000 flowering plant genera, including progress in four major areas:

1) Developing a stringent set of standards to guide the selection of samples included in PAFTOL.
2) Developing an accessible collecting protocol for sampling living plant material to ensure consistent quality and optimal preservation of the samples collected for PAFTOL.
3) Assessing the availability of material in Kew's collections and completing a gap analysis. The assessment showed that of the 8,200 accepted fungal genera, Kew has specimens representing roughly 6,000 of them. And of the ca. 14,000 flowering plant genera, 95% are available across Kew's collections.
4) Harnessing expertise and engagement across Kew through 19 sampling projects that contribute to the overall aims of PAFTOL while also chiming with individual research interests.

In year two, WP1 focused on the following areas:

1. Completing the sampling for the Pilot Project

The plant Pilot Project represents a major milestone for PAFTOL. The Pilot Project requires data to be obtained from all 416 flowering plant families, using samples of 384 genera processed by PAFTOL and combining them with existing data. The resulting DNA sequence data are then processed into a family level tree of life by WP3. In year two WP1 completed the sourcing and DNA extractions for all samples required for the pilot.

2. Forecasting and creating a balanced taxonomic sampling for 2018

In year one PAFTOL compiled a list of all ca. 14,000 flowering plant genera using checklists from RBG Kew and other leading botanical institutions. In the same year, PAFTOL issued a call to Kew researchers for sub-projects that contribute to PAFTOL's sampling goals. In total, 19 proposals were submitted, covering more than 6,400 genera. As a result, samples entering the WP1 workflow come from numerous researchers and a range of Kew sources including the Herbarium, the Millennium Seed bank, the living collections and the DNA bank. Consequently, a major output of WP1 is the ongoing processing and balancing of samples entering the PAFTOL workflow to create a taxonomically balanced dataset. The sampling strategy for 2018/19 is now finished, and it is estimated that between 3,500-4,000 samples will be processed by the end of Year 3, equivalent to 25% of all flowering plant genera.

3. Developing a PAFTOL information system

In order to decide which new samples should enter the PAFTOL workflow, WP1 needs detailed records of all samples that have already been processed, and their success at different stages of the workflow. To facilitate the planned high rate of sample processing, PAFTOL needs a powerful information system (the "PAFTOL database") for storing and querying all sample data. In year one, PAFTOL established many of the basic requirements; in year two, PAFTOL liaised with Kew's IT department to develop a system that meets the needs of all work packages. A beta version is now available and is currently being tested by WP1.

4. Geographic gap analysis of available sequence data

As PAFTOL proceeds to place more and more genera and species in the Tree of Life, it is important to ensure that the samples entering the workflow are balanced not just taxonomically, but also geographically. If proceeding uncritically, well-collected regions of the world will be overrepresented, biasing our understanding of plant evolution. To avoid such bias, WP1 has put major efforts into compiling geographic occurrence data for all plant genera and species, and linking these to available sequence data (including both data generated by PAFTOL and from other sources) to highlight parts of the world with especially low phylogenetic knowledge, which ought to be targeted by future sampling. PAFTOL aims to use this analysis for its sampling as well as publish it as a scientific paper of wider interest.

Phylogenomics - Work Package 2

WP2 is led by senior researcher Dr. Steven Dodsworth, and supported by two research assistants, Niroshini Epitawalage and Grace Brewer. WP2 is responsible for producing the genomic sequence data for plants and fungi, primarily using DNA samples from Kew's collections provided by WP1. The data produced by WP2 feeds into WP3 for bioinformatics analysis and dissemination. In year one WP2 focused on:

1) Procuring equipment and ensuring that Kew's genomics facilities were ready for a large increase in phylogenomic activity.
2) Developing a lab methodology to produce genomic data from DNA samples generated by WP1.
3) Designing the PAFTOL bait kit.
4) Processing samples to generate data for the PAFTOL pilot project.

In year two WP2 focused on the following areas:

1. Testing and evaluating the PAFTOL bait kit

Baits are short molecular (RNA) probes which bind to specific target genes and can be used to "fish out" those targets. In year one, in collaboration with partners at Chicago Botanic Garden, PAFTOL successfully designed and synthesised the baits for use in the project. The PAFTOL bait kit consists of approximately 72,000 individual bait sequences of 120 bp each, which can fish out a maximum of 353 genes. The plant trees of life will then be constructed through comparative analyses of the DNA sequences from these genes.

In year two, WP2 tested and evaluated the bait kit's effectiveness and utility. A publication on the bait sequences and the methods used to develop them is currently in preparation and will be submitted shortly. Concomitant with this, PAFTOL will release the bait sequences and kit via Arbor Biosciences (formerly MYcroarray), who will make it available to the scientific community at low cost, thus increasing the worldwide impact of PAFTOL's work.

2. Completing the PAFTOL pilot – sequencing one sample for all families of flowering plant

WP2 sequenced DNA from 384 samples in order to complete family-level sampling for the pilot project. Throughout the pilot phase WP2 has included samples that vary taxonomically, by genome size, and by sample type (e.g. living sample, herbarium sample, etc.). WP2 has also tested the hybridisation step to the bait sequences and continues to refine procedures for retrieving as much target DNA data as possible. For the 384 sequenced samples, WP2 has recovered target sequences for a median of 200 of the 353 genes from across the entire sample

set, with variability dependent on several factors that are being explored further. A major manuscript is currently in preparation that describes these results.

3. Evaluating potential of the PAFTOL baits in systematic studies between species

The primary application planned for the PAFTOL bait kit is the reconstruction of the tree of life among the flowering plant genera. However, there is huge demand for tractable genomic tools that will permit the relationships among closely related species to be reconstructed. We gathered preliminary datasets to explore the application of the PAFTOL bait kit to species level phylogenetics in six taxonomically diverse groups across flowering plants: *Nymphaea* (water lilies); *Babiana* (iris family); *Basselinia* (palm family); *Combretum* (bushwillow family), *Nepenthes* (pitcher plants) and *Nicotiana* (tobacco). Preliminary results are extremely promising and imply that the impact of the PAFTOL bait kit will be wide reaching.

4. Testing the potential of the PAFTOL baits in DNA barcoding

Related to this, DNA is used extensively as a tool for molecular identification of species, a technique known as DNA barcoding. Barcoding has many potential applications, such as enabling the identification of fragments of unknown organic material and has huge potential in ecology and commerce. Until now, barcoding has been performed using a limited number of genes, which have not been universally effective in plants. The PAFTOL bait kit, however, may function as a "next generation" DNA barcoding technique, which could offer vastly greater discriminating power than existing tools. Preliminary results (in *Nicotiana*) show that very closely related species were successfully discriminated. Full results of this investigation will be available by the end of 2018.

Case study: Recovering DNA sequences from degraded herbarium specimens

Ideally, PAFTOL samples are sourced from samples likely to yield high quality DNA for sequencing, such as Kew's living collections, the millennium seed bank, or the DNA bank. However, 43% of genera are not available from these sources. In such cases, PAFTOL uses material from the herbarium. The Kew herbarium collection is estimated to hold more than 95% of recognised flowering plant genera. However, the quality of the DNA obtained from herbarium specimens varies dramatically due to DNA degradation that occurs naturally during drying of the sample and subsequently over time.

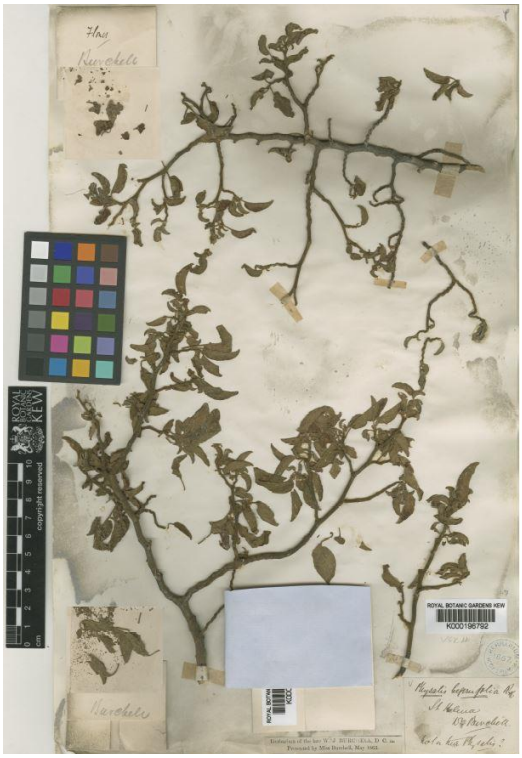


Image 1: The specimen of Boxwood (Mellissia begonifoli) from Kew's Herbarium.

Using our novel protocol, PAFTOL has successfully retrieved excellent DNA sequence data from herbarium specimens, even where samples are extremely old. One of the most outstanding examples so far comes from *Mellissia begonifolia,* commonly named Boxwood, a critically endangered species endemic to St. Helena with a wild population of only 50 individuals. PAFTOL has successfully extracted DNA from a specimen of Boxwood collected by William John Burchell between 1805 and 1810 and generated sequence data of sufficient quality for it to be included in analysis.

Samples from the herbarium that were previously thought to be unsuitable for systematic work because of their age have now proven to be viable

for high throughput DNA sequencing. This will not only enable PAFTOL to source many of the genera needed to complete the project from the herbarium, but will also enable us, for the first time, to unlock the genomic secrets of the herbarium and add value to the historical specimens, some of which are now extinct.

Bioinformatics - Work Package 3

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 WP3, led by senior bioinformatician Dr Jan Kim, is to process the sequencing data generated by WP2 into a comprehensive tree of life for plants. Dynamic inference – i.e. the ability to automatically incorporate new data and update the tree – is a key target. This will result in a constantly relevant piece of scientific infrastructure with wide and continuing utility past the formal end of the PAFTOL project.

In year one, WP3 delivered the following:

1. The procurement of a computer cluster needed to analyse PAFTOL genomic data.
2. The development of a tool, the 'PAFTOL pipeline', to analyse the genomic data produced by WP2.
3. Specifying the requirements for an information system to track the progress of a sample through the work packages.

In year two WP3 focused on the following areas:

1. Developing the PAFTOL analysis pipeline

WP3 has focused on developing and improving its software pipelines to ensure that trees of life are constructed in a robust, flexible and scalable manner. Dynamic inference is a key objective for these pipelines. Once fully established, these pipelines will be made available to the scientific community as a permanent resource and will form part of the legacy of the PAFTOL project.

2. Software developed and enhanced.

The WP3 team has developed a software framework that provides tools for recovering PAFTOL gene sequences 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, the code has been 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 sequencing reads especially with samples that are highly divergent from the known reference sequences.

3. Developing multigene phylogeny inference software

The software for the subsequent step of tree of life inference is at an earlier stage of development and still requires a considerable amount of human interaction. Nonetheless, software for the entire process of inferring multigene trees of life is now complete, and it has been used successfully with the pilot data, and to infer multigene phylogenies for specific taxonomic groups.

The Plant pilot:

The pilot project has been an excellent testing ground for the processes, methodologies, equipment and systems required to process a sample through the three work packages. In completing the pilot the team have been able to identify and refine their techniques and approaches, which they will continue to perfect and develop as the project matures. With this knowledge and expertise in hand PAFTOL is in a strong position to deliver its ambitious aim of completing the Plant and Fungal Trees of Life by 2020.

Of equal importance, the pilot has produced a significant scientific output, a family level tree of all 416 flowering plant families. This tree, along with the methods and analysis used in its creation are due to be submitted for publication in 2018 in multiple papers.

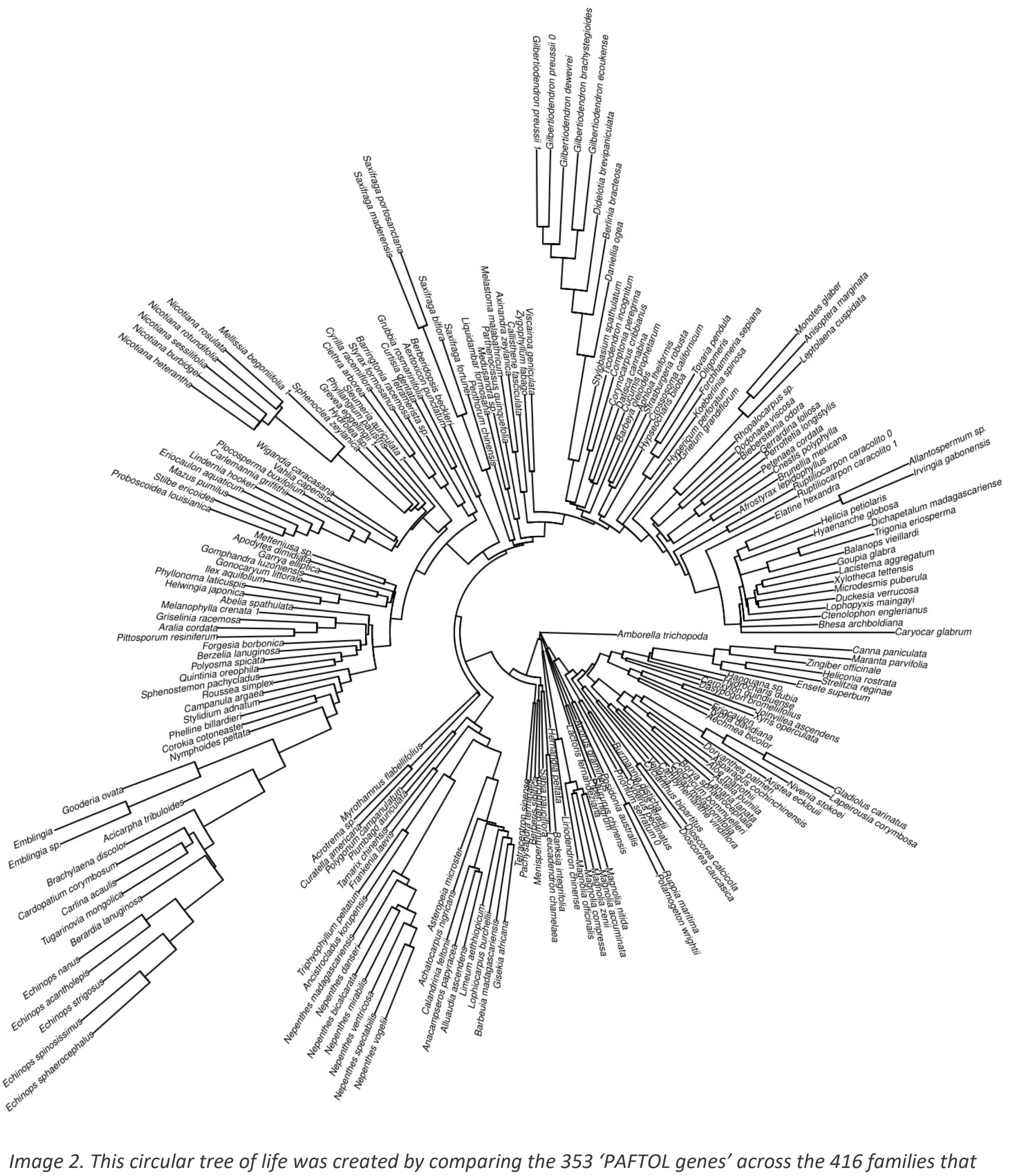


Image 2. This circular tree of life was created by comparing the 353 'PAFTOL genes' across the 416 families that make up the flowering plants.

The Fungal Pilot:

Although the PAFTOL project has thus far primarily focused on constructing the plant tree of life, substantial headway has been made with completing a fungal pilot too.

Fungi belong to one of the largest and most diverse kingdoms of living organisms, with an estimated 1.5-5 million species, from more than 8,200 accepted genera. Fungi play a pivotal role in the functioning of ecosystems; they exhibit a wide variety of life cycles, metabolisms, morphogenesis, and ecologies that have enabled a diversity of other organisms to exploit novel habitats and resources. Despite their enormous ecological importance, the evolutionary relationships between fungal genera are not yet well understood or resolved. For these reasons, a robust fungal phylogeny will greatly enhance our understanding of the history of life.

Until recently, evolutionary relationships within fungi have been inferred mostly from a small number of highly conserved genes common amongst fungi. In comparison to plants, fungi have a relatively small genome size, the average fungal genome size is around 35 megabases (i.e. 50 million letters of genetic code), whereas in plants it is around 6 gigabases. For this reason, and due to the continuing fall in the cost of sequencing, whole-genome sequencing of fungi has become viable.

The PAFTOL fungal pilot aims to sequence whole genomes for a total of 100 taxa, representing a major step forward in our understanding of fungal evolutionary relationships. The pilot is composed of three sub-projects, each of which is focused on the evolutionary relationships within a specific group (Agaricinae, Agaricaceae and Boletaceae).

In year one, a gap analysis was conducted to identify how many good quality fungal samples (those less than 15 years old) were available from Kew's Fungarium. This analysis revealed that of the 8,200 accepted fungal genera, there were 1,500 genera of good enough quality to be sequenced straightaway and a further 4,500 genera that would be tractable with additional effort to obtain a sufficient DNA yield.

In year two, the fungal pilot team focused on the Agaricinae sub project. At the current time, DNA has been extracted and 24 full genome sequences obtained for this group. The next step will be to compile the Agaricinae genomes with publically available genome sequences from 11 species, as well as fungal ITS barcoding data, to place these species in a fungal backbone tree.

In year three the fungal pilot team will obtain genome sequences from the Agaricaceae and Boletaceae subprojects and add these to the fungal tree of life.

Stakeholder and public engagement

During the past year, PAFTOL has continued to connect and engage with key stakeholders relevant to the project. In particular, PAFTOL has interacted closely 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). All of these people are key players in the 1KP (one thousand plant transcriptome project) and have been immensely supportive of PAFTOL's goals.

PAFTOL participated in the founding workshop of the Earth BioGenome Project (EBP) at the Smithsonian Institution in 2016, an ambitious "moon-shot" vision to sequence the genomes of all species of life on Earth. Kew has subsequently participated in the EBP's working group, including in a position paper soon to be published in the *Proceedings of the National Academy of Sciences of the USA*. Kew has also been at the table when EBP was the focus of a day-long workshop at the Wellcome Trust aimed at determining the UK's response to this agenda; William Baker gave a presentation on the project at this workshop. Kew has become a credible player in this arena on account of our ambition stated through the PAFTOL project and is now widely seen as a key counterpart in comparative genomics in the UK.

The PAFTOL team also participated in numerous conferences, seminars and workshops, presenting the project to a wide, international audience. Most significant among these was the International Botanical Congress in Shenzhen, China (July 2017), at which two talks were given on PAFTOL vision and methods, attracting significant attention and buy in from other scientists.

Image 3. PAFTOL staff demonstrating the basics of DNA extraction techniques to visitors from the Global China Philanthropy Forum.

Notably, PAFTOL hosted a workshop on genomic methods in May 2017, which was taught by a team of our collaborators from Chicago Botanic Garden (namely Norm Wickett, Matt Johnson and Elliot Gardner). This was hugely popular and successful, with both internal and external participants.

In October 2017, PAFTOL hosted the inaugural meeting of the PhyloSynth network, which involved numerous eminent scientists from the US and Europe with shared interests in phylogenetic synthesis for the plant tree of life. The group explored ways that the growing volume of genomic data could be integrated with vast existing DNA datasets that are available in public data repositories. The workshop resulted in a vision paper which has just been published in the American Journal of Botany. The network will hold its second meeting at the upcoming Evolution meeting in Montpellier in August 2018.

PAFTOL has also pursued opportunities to share its work with the public through the Kew Science Festival, and Kew's Circle of Benefactors Dinner. The PAFTOL team has also taken the opportunity to engage with other influential individuals, especially visiting VIPs, such as the Global China Philanthropy Forum.

Plans for 2018-2019

General

- Submit a paper to a major scientific journal publishing the PAFTOL baits.
- Submit 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.
- Further stakeholder engagement with the scientific community through participation in at least three international conferences, and the delivery of a PhyloSynth workshop.
- Development of the PAFTOL web presence.
- Refining and finalising the scope of the PAFTOL Explorer.
- Further recruitment to the project team.

Sampling – Work Package 1

- Source and process all samples required to meet the 25% target of 3,500 specimens by March 2019.
- Balance the sampling for the 25% target to ensure that the sampling is taxonomically representative.

Phylogenomics – Work Package 2

- Refine lab processes and techniques, including assessment of current technology and time-limiting steps, as well as factors affecting hybridisation success.
- Establish an outsourcing contract to scale up the production of sequence data.
- Deliver sequence data to meet the target of 25% sampling of flowering plants by March 2019.

Bioinformatics – Work Package 3

- Further improve the bioinformatic pipelines for analysing the sequence data, with the focus shifting from sequence recovery and assembly to subsequent stages of alignment and phylogeny reconstruction.