Centre for Forest Protection

CFP Science Project Internship Diary
Elm Project

Mahya Khakpash

January-March 2026

Royal Botanic Gardens Kew

Motivation for This Internship

After completing my research Master's degree, I decided to continue in multidisciplinary plant science research and aim for a PhD. I am strongly motivated to pursue a career in plant science research.

This internship provided an excellent opportunity to gain experience in bioinformatics and genomic analysis and further develop the skills needed for my future research career.

My Project

Fig. 1. Jodrell Laboratory, Royal Botanic Gardens, Kew, where my internship was based.

My internship project was part of the larger ongoing Elm project and focused on two main objectives. The first was to reconstruct a phylogenetic tree of *Ulmus* L. species by combining two types of molecular data: short-read whole-genome sequencing (WGS) and publicly available restriction-site associated DNA sequencing (RADseq) data (Whittemore et al., 2021). Integrating these datasets allowed us to examine how both marker types represent the phylogenetic relationships among elm species. The second objective was to map the evolution of **Dutch elm disease (DED)** across the elm family.

Research Workflow and Skills Developed

1. Reading and Note-Taking

I reviewed relevant scientific articles to build background knowledge and become familiar with DED.

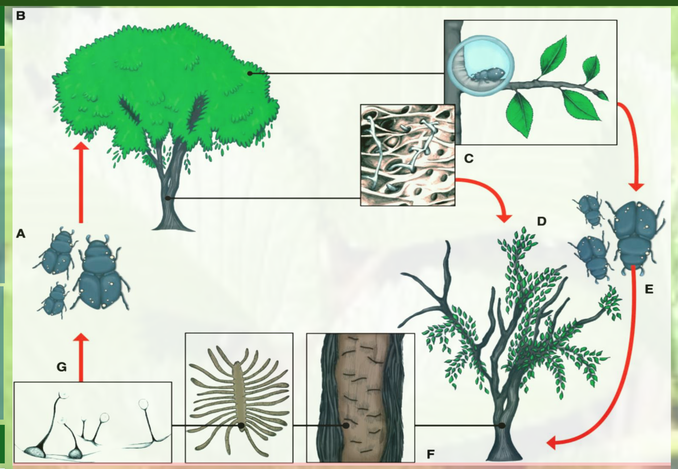
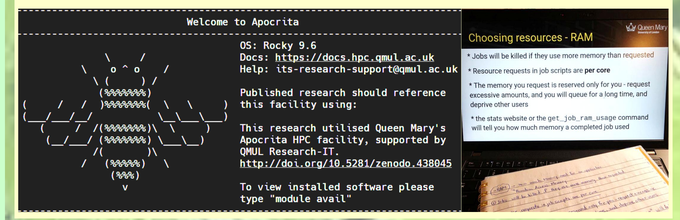


Fig. 2. Disease cycle of DED.
DED is caused by the fungi *Ophiostoma ulmi*, *O. novo-ulmi*, and *O. himal-ulmi* and transmitted by elm bark beetles. (A) Beetles carrying fungal spores feed on healthy elm trees (B), introducing the pathogen into the xylem. (C) The fungus spreads through the vascular system, leading to wilting and tree death (D). Dead or weakened trees attract beetles for reproduction (E–F), where the fungus grows within galleries and produces reproductive structures (G). Spores attach to emerging beetles and continue the transmission cycle. Source: Comeau et al. (2015).

2. Linux Command Line and Apocrita HPC Cluster

One of the most valuable skills I developed during this internship was learning to use the Linux command line and the Apocrita HPC cluster at QMUL. This allowed me to run large-scale genomic analyses and manage computational jobs efficiently.



3. Quality Control and Preprocessing

I learned to use several bioinformatics tools for preparing sequencing data. **FastQC** was used to assess the quality of raw DNA sequencing reads (FASTQ files), providing information on sequence quality scores, GC content, and potential technical issues.
Trimming tools were then used to remove low-quality bases and adapter sequences from the reads, improving the overall data quality before downstream analyses such as RADseq assembly and phylogenetic reconstruction.

4. Initial Data Analysis: ipyrad and IQ-TREE

I practised RADseq data analysis using **ipyrad**, a bioinformatics pipeline used for assembling and processing RADseq data for population genetics and phylogenetic studies. Ipyrad clusters sequencing reads into loci, filters low-quality data, and generates datasets suitable for downstream analyses. I first practised this workflow using an example dataset and then learned to use **IQ-TREE** to reconstruct a phylogenetic tree.

5. Main Data Analysis: ipyrad and RAxML

I ran ipyrad to assemble loci from combined WGS and RADseq datasets and generate a dataset for phylogenetic analysis. The dataset consisted of 203 samples (108 WGS and 95 RADseq). The final assembly retained 4,507 loci, producing an alignment of 388,294 bp containing 60,597 SNPs across 203 samples for downstream phylogenetic inference.

The resulting dataset was then analysed using **RAxML-NG** under the GTR+G model with 200 bootstrap replicates and 20 parsimony starting trees to reconstruct the phylogenetic tree.

6. Determining DED Resistance Levels

I searched the literature to collect information on the resistance levels of DED for each *Ulmus* species and the outgroup genera included in the dataset, such as *Planera* J.F.Gmel., *Zelkova* Spach, and *Hemiptelea* Planch. Resistance levels were coded as **0 = resistant, 1 = susceptible, and 2 = unknown** to create a trait dataset for evolutionary analysis.

7. Character Evolution Analysis

Character evolution analysis to map Dutch elm disease (DED) resistance across the elm phylogenetic tree using tools such as **Mesquite or BayesTraits**.

Additional Benefits of the Internship

- The research environment at the Royal Botanic Gardens, Kew was supportive and collaborative. Regular meetings with my line manager and the lab team helped me make steady progress during this short-term internship. My line manager supported me throughout the project, guiding me in learning new bioinformatics techniques and helping me gain confidence to work more independently.
- These meetings not only helped me improve my communication skills, but also increased my confidence in speaking and discussing research in groups.
- In addition, I used the internship training budget to enrol in an online course on the **R language** organised by the University of Liverpool, gaining valuable data analysis skills for my future PhD research.

References

- Whittemore et al., 2021. DOI: https://doi.org/10.1600/036364421X16312068417039
- Comeau et al., 2015. DOI: https://doi.org/10.1093/gbe/evu281
- Eaton & Overcast, 2020. DOI: https://doi.org/10.1093/bioinformatics/btz966
- Kozlov et al., 2019. DOI: https://doi.org/10.1093/bioinformatics/btz305
- Background photo: *Ulmus minor* Mill., Andrea Moro/Dryades, EUFORGEN

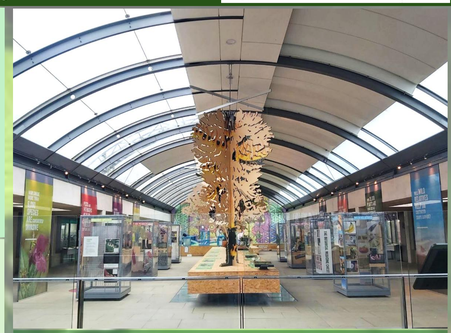


Fig. 3. Millennium Seed Bank Tour, Wakehurst
Visiting the Millennium Seed Bank and meeting other CFP interns provided valuable insights into global seed conservation and research collaboration. (February 2026)

Fig. 4. MSB, Wakehurst (February 2026)

Fig. 5. Kew Herbarium Tour (March 2026)

Fig. 6. Early spring at RBG Kew (March 2026)