Patricia Chan

Phylogenetics Methods Final Project

Botany 563

**Abstract:** Darwinia (Myrtaceae) contains (~65) species with a wide array of specialized floral structures and pollinators ranging from thynnid wasps and nectar-feeding birds to marsupials. This genus is known to have diversified rapidly over small spatial scales, with some species limited to remarkably narrow geographic regions. Previous phylogenetic studies in this group Disagreement between nuclear and organellar genomes are common, especially at shallow evolutionary scales. This case study examines a small subset of species across the flowering plant genus *Darwinia,* comparing maximum likelihood phylogenies generated using plastid and nuclear assemblies resulting from the same Illumina sequencing data. This demonstration acts as a small scale test of approaches that will inform future phylogenetic analyses of *Darwinia* and subsequent exploration of the drivers of plant diversification. While results were not successfully obtained, we suggest further analyses of plastomes in conjunction with nuclear data to further elucidate evolutionary relationships within Darwinia.

**A screenshot of a cell phone

Description automatically generatedA close up of a flower

Description automatically generatedIntroduction:**

*Darwinia* (Myrtaceae) is flowering plant genus composed of ca. 65 species1 that is endemic to the Mediterranean climates of the Southwest Australian Floristic Region, a global biodiversity hotspot with especially narrow plant species ranges2. This genus’ narrow distribution on isolated mountaintops, unique seed dispersal strategies and high diversity in floral morphology (Fig. 1) make it a model system to examine the relative importance of multiple drivers of plant speciation. As seen in figure 1, Darwinia spp. display a broad array of highly specialized floral structures associated with unique ecological relationships,

Preliminary phylogenetic analysis conducted with nuclear ITS and ETS sequences by Dr. Mathew Barrett (pers comm.) suggest that *Darwinia* as currently described, is polyphyletic. In order to resolve this polyphyly, he suggests nesting *Darwinia* within the nearby genus *Verticordia*, and splitting the genus, isolating the eastern and southwestern Australian clades. The relationships within *Darwinia* based off ITS and ETS data are largely unresolved or has low bootstrap support. One of the goals of my dissertation research is to build a more robust and better resolved phylogeny of the southwest Australian clade of *Darwinia* using next-generation sequencing data and target sequence capture approaches.

Text

Description automatically generatedAs a proof of concept, we have collected and sequenced 11 species distributed across eastern Australian clade, focusing on the most recently diverging crown group which has historically lacked support for its relationships. The raw Illumina sequencing reads featured in this study have been used to produce a RaxML3 phylogeny based on concatenated nuclear data (Fig. 2). This preliminary tree is promising in that the bootstrap values (90-100) indicate higher confidence in species relationships, resolving previous polytomies and areas of low confidence seen in earlier phylogenies created with Sanger sequencing data.

In this study, I planned to align the same Illumina reads and construct whole chloroplast genomes (hereby referred to as plastomes) using a reference-based assembly pipeline. The whole plastomes would be used to construct a maximum likelihood tree to compare against the existing nuclear phylogeny (fig 2) generated using the same samples. I expect to see some disagreement between the nuclear and plastome trees which is typically seen at shallow evolutionary levels due to the different evolutionary forces acting upon nuclear and organellar DNA4,5.

These data represent only a fraction of the genus *Darwinia*; the main purpose of this study is to demonstrate use of this phylogenetic pipeline, which will later be scaled up to construct a more complete phylogeny of all ca. 65 *Darwinia* spp. Trees generated using this pipeline on the full dataset can then be used to estimate speciation rates, map floral morphometric traits, reconstruct historical biogeographic patterns, and infer ancestral traits.

**Materials and Methods:**

*Data Collection and Sequencing:*

Dr. Matthew Barrett collected silica-dried leaf tissue of one individual each of two (2) *Verticordia* and eleven (11) *Darwinia* species from which I extracted total genomic DNAs using Qiagen DNeasy kits 6. Aliquots of DNA were sent to our collaborators Emily and Alan Lemmon at Center for Anchored Phylogenomics7 for Illumina library preparation, hybrid DNA enrichment for ca. 400 nuclear loci, and next-generation sequencing. The resulting sequencing data was returned in the form of 4 .fastq files per sample representing two sequencing lanes of paired-end reads.

*Sequence Trimming and Quality Control:*

Software choices in assembly were informed by lab members who have constructed plastomes using similar datasets. I conducted the plastome assembly pipeline (see Darwinia\_plastomes.bwa.sh) using University of Wisconsin-Madison’s Center for High Throughput Computing (CHTC) to allow parallel processing of samples for a scalable process that can be conducted in a reasonable amount of time. Subsequent steps were attempted on a local machine for simplicity; however, future work will be conducted remotely using the CIPRES supercomputer portal8 or CHTC to accommodate larger scale analyses. S

I processed the resulting raw Illumina sequencing reads using Trimmomatic9, selected for the tool’s relative flexibility and efficiency. Trimmomatic trimmed removed custom Illumina sequencing adaptors with a simple clip threshold of [2:30:10], allowing 2 mismatches within seeds, clipping paired-end reads with a minimum score of 30, and single-end reads with a minimum score of 10. Poor-quality read ends were trimmind with a sliding window of 5 bp, clipping reads under a Phred score of 20. Trimmed reads were visualized and checked using FastQC10 before and after trimming to verify the removal of adaptor content and poorly sequenced regions. Reads were primarily screened in FastQC for per base sequence quality, per tile sequence quality, per sequence GC content, and adaptor content.

*Plastome Assembly:*

I aligned trimmed reads to a reference genome sequence using Burrows-Wheeler Aligner, BWA11 for a reference-based assembly. BWA was selected for its reliability and popularity in phylogenetics pipelines lending to large support communities. *Eucalyptus albens* was selected as a reference for its phylogenetic closeness to *Darwinia* and genomic data access via the NCBI Genome database12. I then used SAMtools13 to sort the .bam files, remove PCR duplicates, and phase SNPs from the resulting assemblies. To check the quality of the plastomes assemblies, I used SAMtools and BCFtools to produce summary statistics on read depth and coverage as well as a QC report of alignments on the .bam file. bcftools mpileup and bcftools call were used to call variants from the sorted and filtered .bam file. variants are then to be applied to the reference sequence to generate a consensus sequence for the assembled sample plastome. Unfortunately, bcftools consensus was unable to successfully read and apply variants from the .vcf.gz files generated with mpileup, and returned consensus.fasta files each identical to the reference. The following methods describe my planned pipeline for the assembled plastomes, had I successfully generated them.

*Multiple Sequence Alignment and Tree Building:*

After concatenating all samples into a multifasta file, I would visualize the sequences side-by-side using the graphical user interface Unipro UGENE14. UGENE allows for easy visualization, provides guided workflow templates and a toolkit to access various programs to handle genetic data including those for multiple sequence alignment and tree building. I would use MAFFT15 to align the assembled plastomes using default parameters. MAFFT was selected due to its accuracy as a progressive-iterative aligner, and for its multi-threaded approach which optimizes speed, even with large datasets16. It operates on the assumption of homology of input taxa and thus possibility of alignment.

I would employ IQ-Tree17 to generate maximum likelihood trees from the aligned plastomes which will be comparable to the existing nuclear RaxML trees. Through IQ-Tree, I would use ModelFinder18 to test and select the appropriate models for nucleotide substitution. IQ-Tree was selected for its ease of use, and built-in model selection and ultrafast bootstrap approximation features that run 10-100x faster than comparable programs19. IQ-TREE operates on the assumptions of homogeneity and stationarity which can be tested within the program. Resulting trees with support values would be viewed and manipulated using FigTree.

*Generation of hypothetical data:*

Due to the obstacles encountered in the assembly process, a ML phylogeny was not successfully generated using the methods described above. For the purpose of this report, I fabricated a tree to illustrate theoretical results and how I would interpret them within the context of this project. This tree was manually generated in R from a parenthetical text string.

**Diagram

Description automatically generatedResults:**

Our resulting phylogeny highly resembles the nuclear tree generated using the same samples, with several notable differences in topology (see fig 3). The crown clade represented *by D. polycephala*, *D. sp, Mt Ragged* and *D. sp. Gibson* is retained and largely highly supported with node bootstrap values ranging from 95-100. While *D. masonii* was recovered as sister to this crown group in the nuclear tree, *D. oldfieldii* is seen as sister in the plastome tree. Interestingly, the deeper nodes with lower support values in the nuclear tree have also been collapsed into a polytomy with *D. sp. Dryandra*. As expected, the *Verticordia* species remain grouped together and have emerged as an outgroup to *Darwinia*. Bootstrap values across this tree show lower support for nodes than those of the nuclear tree

Fig. 3: Hypothetical maximum likelihood plastome tree with bootstrap values illustrating possible output from IQ-Tree. Discordance in topology between the nuclear (Fig 2) and plastome trees highlighted in orange.

Promisingly, both nuclear and plastome trees demonstrated high support in crown groups which have historically been challenging due to the recent rapid radiations in the genus. This shows potential for the usage of whole plastome data in future genus-wide analyses in *Darwinia*, in contrast to previous methods examining a limited number of plastic loci. While plastomes generated with these methods may not replace the usage of nuclear genomic data, they may be used in tandem to increase support and resolution of inferred evolutionary relationships.

**Discussion:**

This study served as a small-scale model to demonstrate the potential for constructing and analyzing plastomes from byproducts of next-generation sequencing focused on nuclear loci. While results have yet to be yielded from this example, these approaches applied to similar data indicate that these methods can be used to complement nuclear data to better infer evolutionary relationships in recently diverging groups such as those within *Darwinia*.

There were clear limitations to the approaches executed, made partially evident by my lack of results. BCFTools and SAMtools provide clear documentation. However, they would benefit from addressing common pitfalls and sustaining a more robust online communities. This is supported by the volume of forum unresolved threads on Biostar troubeshooting the command “bcftools consensus”. Errors aside, these tools provide a wide array of options for applicability over many data types

The Burrow-Wheeler Aligner has served as a staple in phylogenetics pipelines since its inception, and has since been iterated upon, producing improved versions. In applying this pipeline to larger datasets of a greater number of samples or longer sequences, I would consider utilizing bigBWA20 or BWA2 to optimize time efficiency. Going forward, I also would consider alternative pipelines to plastome assembly, including GetOrganelle21 and FastPlast which both provide streamlined toolkits for rapid and automated de novo assemblies. A *de novo* approach to plastome assembly could be beneficial in assembling plastomes unbiased by a reference outside of the focus group. However, *de novo* approaches may not be viable given this dataset if there are not enough overlapping contigs to produce sufficient coverage. Other case studies have suggested that hybrid approaches using both *de novo* and reference-based assemblies may produce the best results22.

Future directions with this dataset include comparing different inference methods between Bayesian and maximum likelihood. Bayesian inference would be conducted using MrBayes23 and an additional maximum likelihood tree would be conducted using RaxML3. I would then calibrate the resulting plastome tree against time with BEAST24 using primary and secondary divergence dates within the order Myrtales, obtained by Berger et al25. Ultimately, I hope to employ these methods on a full dataset representing all species across Darwinia to provide a backbone for understanding the divergence and convergence evident in the group, and relating its diversification patterns back to evolutionary and ecological drivers

**References:**

1. Keighery, G. J. Six new and rare species of Darwinia (Myrtaceae) from Western Australia. https://www.cabdirect.org/cabdirect/abstract/20093174230 (2009).

2. Cook, L. G., Hardy, N. B. & Crisp, M. D. Three explanations for biodiversity hotspots: small range size, geographical overlap and time for species accumulation. An Australian case study. *New Phytol.* **207**, 390–400 (2015).

3. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

4. Stull, G. W., Soltis, P. S., Soltis, D. E., Gitzendanner, M. A. & Smith, S. A. Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. *Am. J. Bot.* **107**, 790–805 (2020).

5. Zhang, X. *et al.* Plastome phylogenomic study of Gentianeae (Gentianaceae): widespread gene tree discordance and its association with evolutionary rate heterogeneity of plastid genes. *BMC Plant Biol.* **20**, 340 (2020).

6. Drábková, L., Kirschner, J. & Vlĉek, Ĉ. Comparison of seven DNA extraction and amplification protocols in historical herbarium specimens of juncaceae. *Plant Mol. Biol. Report.* **20**, 161–175 (2002).

7. Mitchell, N., Lewis, P. O., Lemmon, E. M., Lemmon, A. R. & Holsinger, K. E. Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L. *Am. J. Bot.* **104**, 102–115 (2017).

8. The CIPRES science gateway | Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery. https://dl.acm.org/doi/abs/10.1145/2016741.2016785?casa\_token=4HKLzU0rlboAAAAA%3AfETJObWsnFiKh4ftDA2nXLLa4z3FLMiYo0ZBLLqO6HHY-ItRVZtKhDEXE3U9OYoANY1fUMvBcGnw.

9. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

10. Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research* **7**, 1338 (2018).

11. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).

12. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).

13. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

14. Okonechnikov, K., Golosova, O., Fursov, M., & the UGENE team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167 (2012).

15. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).

16. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).

17. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

18. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

19. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast Approximation for Phylogenetic Bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).

20. Abuín, J. M., Pichel, J. C., Pena, T. F. & Amigo, J. BigBWA: approaching the Burrows–Wheeler aligner to Big Data technologies. *Bioinformatics* btv506 (2015) doi:10.1093/bioinformatics/btv506.

21. Jin, J.-J. *et al.* GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 241 (2020).

22. Marchant, A. *et al.* Comparing de novo and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug Rhodnius prolixus. *Insect Biochem. Mol. Biol.* **69**, 25–33 (2016).

23. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).

24. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).

25. Berger, J., Palta, J. & Vadez, V. Review: An integrated framework for crop adaptation to dry environments: Responses to transient and terminal drought. *Plant Sci.* **253**, 58–67 (2016).