Harnessing Chloroplast Diversity for High-Throughput DNA Barcoding

Cornelia M Hooper[1], Julian Tonti-Filippini[1], Joao Melanda Filipe[2], Zhong Xiao[1], Ian R Castleden[1], Ian D Small[1], Rachel Binks[2], Benjamin Anderson[2,3], Stephen Van Leeuwen[3,4], Kevin Thiele[3], Paul Nevill[4,5] and Margaret Byrne[2]

[1]The ARC Centre of Excellence in Plant Energy Biology, School of Molecular Science, The University of Western Australia, Crawley, WA 6009, Australia
[2]Biodiversity and Conservation Science, Department of Biodiversity, Conservation and Attractions, Kensington, WA 6151, Australia
[3]Western Australian Herbarium, Biodiversity and Conservation Science, Department of Biodiversity, Conservation and Attractions, Kensington, WA 6151, Australia
[4]School of Molecular and Life Sciences, Curtin University, Bentley, WA 6845, Australia
[5]ARC Centre for Mine Site Restoration, School of Molecular and Life Sciences, Curtin University, Bentley, WA 6102, Australia

Surveys of biodiversity hotspots such as the Pilbara region (>1,500 species) in Western Australia generally require multiple visits by expert botanists to remote areas under challenging conditions. Chloroplast DNA barcoding offers a potentially cost-effective route to rapid and accurate species identification from any plant material collected at any time of the year. Comprehensive chloroplast DNA catalogues are an essential basis for such approaches and are also valuable for their insights into chloroplast genome diversity.

Over 1000 samples from Australian native species endemic to the Pilbara and some introduced species were sourced from the WA Herbarium. DNA was extracted, skim-sequenced using Illumina sequencing and chloroplast genomes assembled with GetOrganelle or NOVOplasty. We developed Chloë (https://chloe.plastid.org/), for accurate high-throughput annotation of chloroplast genomes based on a small set of reference genomes that we manually annotated based on RNA-seq data. Our PilbSeq (https://pilbseq.dbca.wa.gov.au) data set now includes 993 plant species found in the Pilbara. We have developed novel species recognition software that uses taxon-specific kmers in the chloroplast genome to identify species present in sequencing reads from mixed samples, making high-throughput, cheap (~$1 per plant) and accurate plant species identification achievable.