RADseq as a valuable tool for plants with large genomes-a case study in cycads

RADseq as a valuable tool for plants with large genomes—a case study in cycads

James A. R. Clugston[1,2], Gregory J. Kenicer[2], Richard Milne[1], Isaac Overcast[3] Trevor C. Wilson[5] and Nathalie S. Nagalingum[4]

1. The University of Edinburgh, School of Biological Sciences, The Kings Building, West Mains Road, Edinburgh, Scotland, EH9 3JN, United Kingdom
2. Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, Scotland, EH35LR, United Kingdom
3. The Graduate Center of The City University of New York, 217 E 42nd St, New York, NY 10017, USA
4. California Academy of Sciences, 55 Music Concourse Dr, San Francisco, CA 94118, USA
5. Royal Botanic Gardens and Domain Trust, Mrs Macquaries Rd, Sydney NSW 2000

Corresponding authors: James A. R. Clugston jclugston@rbge.ac.uk and Nathalie S. Nagalingum nnagalingum@calacademy.org

Keywords: RADseq, cycads, illumina sequencing, large genomes.

Running title: RADseq a tool for plants with large genomes

Abstract

Full genome sequencing of organisms with large and complex genomes is intractable and cost ineffective under most research budgets. Cycads (Cycadales) represent one of the oldest lineages of the extant seed plants and, partly due to their age, have incredibly large genomes up to ~60Gbp. Restriction site associated DNA sequencing (RADseq) offers an approach to find genome-wide informative markers and has proven to be effective with both model and non-model organisms. We tested the application of RADseq using ezRAD across all ten genera of the Cycadales including an example dataset of *Cycas calcicola* representing 72 samples from natural populations. Using previously available plastid and mitochondrial genomes as references, reads were mapped recovering plastid and mitochondrial genome regions and nuclear markers for all of the genera. De novo assembly generated up to 138,407 high-depth clusters and up to 1,705 phylogenetically informative loci for the genera, and 4,421 loci for the example assembly of *C. calcicola*. The number of loci recovered by de novo assembly were lower than previous RADseq studies, yet still sufficient for downstream analysis. However, the number of markers could be increased by relaxing our assembly parameters, especially for the *C. calcicola* dataset. Our results demonstrate the successful application of RADseq across the Cycadales to generate a large number of markers for all genomic compartments, despite the large number of plastids present in a typical plant cell. Our modified protocol was adapted to be applied to cycads and other organisms with large genomes to yield many informative genome-wide markers.

Introduction

The size of an organism's genome greatly affects the cost of sequencing its genome, which in turn affects the number of organisms for which genomic data is available (Andrews et al., 2016). Large genomes are caused by numerous factors such as tandem repeats, pseudogenes, paralogs, polyploidy or a combination of these factors (Guan et al., 2016). Plant genome sizes are highly plastic (Pellicer et al., 2018), ranging from 13.2 Megabase pairs (Mbp) in the genome of *Ostreococcus lucimarinus*, to over 149 Gigabase pairs (Gbp) in the octoploid *Paris japonica* (Pellicer, Fay and Leitch, 2010). As a result of whole genome duplication, gymnosperm genomes are generally larger than those found in many angiosperms, ranging from ~8 Gbp in *Microstrobus* to ~72 Gbp in *Pinus* and *Ceratozamia* (Zonneveld, 2012; Zonneveld and Lindstrom, 2016; Scott et al., 2016; Roodt et al., 2017). Typically, as a result of polyploidy, the on-average large genome size is caused by an inefficiency of gymnosperms at eliminating repeat amplifications in the genome (Pellicer et al., 2018).

Next generation sequencing (NGS) permits sequencing large stretches of a genome to produce DNA sequence data in the Gbp range at relatively low cost. Full genome sequencing may be the best approach for finding informative markers that assist investigating the evolutionary history of a species (Andrews et al., 2016). However, large and complex genomes present problems of cost for existing NGS approaches (Alexeyenko et al., 2014). Further issues include generating enough repeat reads to account for over-representation of highly repeated elements in the genome (Catchen et al., 2017). Additionally, de novo assembly of larger genomes becomes problematic because of repeated elements, making effective repeatability of an assembly difficult (Meyers, Scalabrin and Morgante, 2004).

Restriction-site associated DNA sequencing (RADseq), uses restriction enzymes to reduce the proportion of the genome sequenced by cutting DNA into smaller fragments, and a subset of these fragments (typically between 200-600 bp) is then selected for sequencing (Davey and Blaxter, 2010). Thus, RADseq allows the sequencing of a reduced representation of the genome yet still at a deep level of sequence coverage, especially near specific restriction sites, therefore only a portion of the genome is sequenced (Andrews et al. 2016). Compared to many NGS methods such as shotgun and whole genome sequencing, RADseq is considered quick and economical under most research budgets (Peterson et al., 2012; Toonen et al., 2013).

RADseq has offered new avenues for phylogenetics and population genomics (Table 1) because it does not require the use of a reference genome (Andrews and Luikart, 2014), and has proven to be very effective for population genotyping by identifying thousands of polymorphisms (Mastretta-Yanes, et al., 2015). These polymorphisms include both neutral and non-neutral markers, that potentially reflect a large portion of a taxon's genome which are involved in natural selection and mutation (Narum et al., 2013). RADseq has been applied in population genetics across a range of model plants, such as *Oryza* and *Carex*, as well as non-model plants including *Senecio, Betula, Sisymbrium*, *Mimulus, Passiflora, Psychotria* and *Mangifera* (Roda et al., 2013, Vandepitte et al., 2013, Wang et al., 2013, Guo et al., 2014, Twyford and Friedman, 2015 and Massatti, Reznicek and Knowles, 2016, Nazareno et al., 2018, Warschefsky and von Wettberg, 2019). It has also been used, to a lesser extent, in plant phylogenetics for *Pedicularis*, *Diospyros*, *Quercus*, *Viburnum*, and *Diuris* (Eaton and Ree, 2013, Eaton et al., 2015, Paun et al., 2015, Eaton et al., 2016 and Ahrens et al., 2017).

Currently published fully-sequenced plastome and mitochondrial genomes for the cycads are few, yet this number already appears to provide sufficient evidence to invest in alternative sequencing methods of genomic DNA, such as that of RADseq. Of the ten genera of cycads, eight - *Ceratozamia, Cycas, Dioon, Encephalartos, Macrozamia, Lepidozamia, Stangeria,* and *Zamia* - have documented plastomes (Wu et al., 2007 and Wu

and Chaw, 2015). Yet a comparison of high GC-biased substitutions, gene conversion, and low sequence variability between both theirs and other published gymnosperm plastomes (e.g. *Pinus thunbergii*, *Abies koreana* and *Araucaria* spp.) indicates that the plastid is not an optimal source of variable markers that are useful for population genetics or phylogenetic studies (Tsudzuki et al., 1992, Wu et al., 2007, Jansen et al., 2011, Ruhsam et al., 2015, Yi et al., 2015, Yang et al., 2016 and Zhou et al., 2016). As of yet, the only full mitochondrial genome that has been sequenced is that of *Cycas* (Wu et al. 2007). Compared to published mitochondrial genomes of the closest allies of Cycads (*Ginkgo biloba* and *Welwitschia mirabilis*), only a few number unique and polymorphic sites were found (Guo et al., 2016), which supports that this genomic compartment is equally uninformative as the plastome.

In order to test the effectiveness of RADseq for taxa with large genomes, we used a RADseq technique across a cohort of samples representing ten known cycad genera (Cycadales). We chose cycads because they have particularly large genomes, ranging from ~25-30 Gbp in *Cycas* L. to ~72 Gbp in *Ceratozamia* (Zonneveld, 2012), which appears to be the result of many tandem repeats, pseudogenes, paralogs, and possibly whole genome duplication (Roodt et al., 2017). In addition to having on-average larger genomes, we also chose cycads because there is need for better methods to find more data-rich sequences for the purposes of systematic and population genomic studies. Therefore, forming part of our larger conservation genomics study targeting cycads, we developed a RADseq protocol that is based on a modification of the ezRAD protocol (Toonen et al., 2013). ezRAD differs from other RADseq approaches as it uses a commercially available library preparation kit and does not require specific restriction enzymes to ligate adapters to cut sites (Andrews et al., 2016). Another advantage of ezRAD when compared to other RADseq protocols is that it requires lower initial setup preparation and costs (Andrews et al., 2014).

The aim of the larger project is to understand the evolution and genetic diversity of wild *Cycas* populations. As a proof of concept, we tested our RADseq approach across all cycad genera. This study aimed to: (1) demonstrate that RADseq can be successfully applied to organisms with large, repetitive genomes, such as cycads, (2) generate a sufficient number of loci using de novo assembly for phylogenetic and population genetic analyses, and (3) develop an effective method that can be used for genome skimming. Ultimately, our goal was to demonstrate the effectiveness of RADseq across large and complex genomes to allow others to follow this protocol.

Materials and methods

Sampling strategy. Freshly collected silica-dried leaf material was sampled for all of the ten genera representing 13 species in the Cycadales, from both families—Cycadaceae and Zamiaceae (Table 2). Cycadaceae leaf samples were taken from *Cycas taitungensis* at the living collection of the Royal Botanic Garden and Domain Trust, NSW Australia (RBGS), and samples of *C. armstrongii, C. maconochiei,* and *C. calcicola* were collected from wild plants in the Northern Territory, Australia. For Zamiaceae, *Bowenia spectablis*, *Ceratozamia kuesteriana*, *Dioon mejiae*, *Encephalartos lebomboensis*, *Lepidozamia peroffskyana*, *Macrozamia johnsonii*, *Microcycas calocoma*, *Stangeria eriopus,* and *Zamia integrifolia* samples were collected from the living collection of the RBGS (Table 2).

Additionally, to test the utility of RADseq at population level, samples were collected from 60 individuals of *Cycas calcicola* from natural populations in the Northern Territory, Australia (Appendix I). The samples included three populations from the Litchfield National Park and three populations in the Katherine region—each population consisted of ten individuals of varying ages. In addition, a further 13 samples were sourced from cultivated ex-situ collections of George Brown Darwin Botanic Garden (Darwin, Northern Territory, Australia) and Montgomery Botanical Centre (Miami, Florida, USA).

DNA extraction and quantification. Approximately 0.05 g of silica-dried leaf samples were

ground to a fine powder using a TissueLyser (Qiagen Inc., Venlo, the Netherlands). When present in large amounts, trichomes were removed to improve extraction quality (specifically in *Cycas calcicola*). High molecular weight genomic DNA was extracted using a DNeasy Plant DNA Extraction Mini Kit (3.0 BR DNA assay; Qiagen, Hilden, Germany). Genomic DNA was inspected using a 2% agarose gel to check for the presence of DNA and impurities. A Qubit fluorometer (3.0 BR DNA assay; Invitrogen, Life Technologies, Carlsbad, CA, USA) was then used to determine the quantity (µg/mL) of the extracted DNA for each sample. The target concentration for samples was (≥) 17 µg/mL; samples that yielded less then this amount was either re-extracted or concentrated using a 1:1 ratio of Agencourt AMPure XP magnetic purification beads to sample volume (Beckman Coulter, Inc) by combining multiple extractions (For more detailed laboratory methods, please see supplementary data Appendix II).

DNA normalization and double digest reaction. First, genomic DNA was normalized to a concentration of 500 ng in 42 µL total volume (0.01 µg/mL) using a QIAgility liquid handling robot (Qiagen Inc., Venlo, the Netherlands). Second, using the QIAgility, 5 µL of NEB 10x CutSmart buffer and 1 µL of Bovine Serum Albumin (BSA; to help stabilize the enzyme digestion) was added to each well and mixed briefly for five seconds using a plate mixer (although these steps were performed using a liquid handling robot, they can be performed manually). This mix was stored at 4°C for a minimum of 5 hours—our tests showed that this helps to reduce the effect of DNA methylation, improving the cutting action of the restriction enzymes. Next, double digest reactions were set up using 1 µL of each EcoR1-HF and Mse1 restriction enzymes, mixed by pipetting manually. Reactions were run in a thermocycler for 3 hrs at 37°C with a final 20 min deactivation step at 65°C. Using 2% agarose gel, samples were checked for a smear to indicate the quality of digestion. Lastly, double digest reactions were cleaned using 1.8:1.0 ratio of AMPure XP beads to sample (90 µL of AMPure XP beads to 50 µL of digested DNA) and quantified using a Qubit high sensitivity kit (3.0 HS DNA assay; Invitrogen, Life Technologies, Carlsbad, CA, USA).

Library preparation. RADseq libraries were prepared following the ezRAD protocol (Toonen et al., 2013) in which we tested two different Illumina (Illumina Inc., CA, USA) library preparation kits: firstly, an Illumina TruSeq PCR-Free high throughput dual index kit and secondly, an Illumina TruSeq nano high throughput dual index kit (PCR-based, FC-121-4003). Our initial aim was to use the PCR-Free kit to help reduce the probability of PCR amplification bias. However, after multiple attempts the PCR-Free kit resulted in poor final yields when quantified using qPCR, and after multiple troubleshooting steps, it was deemed unfit for our target group (cycads). However, the Illumina TruSeq nano kit proved to be effective when the input of genomic DNA was increased by 5x the recommended input, i.e., from 100ng to 500ng, due to the amount of DNA which is lost during clean-up and size selection. We followed the ezRAD protocol v3 using half of the recommended volumes of an Illumina TruSeq kit to save costs (Toonen et al., 2013).

Several quality control checks were carried out during library preparation on a select number of samples (16-24 samples) using a high performance LabChip and a Qubit fluorometer; more specifically, DNA size and quantity (µg/mL) were checked after digestion and after size selection. During the final step of library preparation, we modified the ezRAD protocol in the final bead clean, using a 0.8:1 ratio of AMPure XP beads to sample for the removal of excess adapters observed using a LabChip. Final Illumina libraries were validated using a LabChip, cleaned using a 0.9:1 ratio of AMPure XP beads to sample, and quantified using a Qubit high sensitivity kit (3.0 HS DNA assay; Invitrogen, Life Technologies, Carlsbad, CA, USA). Final libraries were normalized to 10 nM and pooled for sequencing. For more detailed laboratory methods, please see supplementary data (Appendix 1).

Sequencing. We aimed to capture around 1 gigabyte (Gb) of sequence data per sample (in a run of 95 libraries) to account for overrepresentation of the plastid genome, and to capture as much of the nuclear genome as possible. Genomic sequencing was carried out using an Illumina NextSeq 500 with 150 bp paired-end high throughput (HT) on a single flow cell. The

NextSeq 500 HT run can capture up to 120Gb of sequencing data, thereby allowing for our sequencing target of one Gb per sample. The sequencing run was also spiked with 20% PhiX sequencing control V3 (Illumina) to account for low sequence diversity caused by the identical enzymatic digestion cut sites in the ezRAD protocol.

Bioinformatics

Quality control and filtering of sequence reads. The NextSeq 500 generated four fastq files for forward and reverse reads (eight files per sample). The four forward fastq files were concatenated into a single forward fastq file and similarly a single reverse file was created, as required for the downstream RADseq assembly. The concatenated forward and reverse fastq files were screened for quality using PRINSEQ v0.20.4 (Schmieder and Edwards, 2011). PRINSEQ allowed the detection of falloff in read quality for a range of samples from each population. The reads were trimmed using Trimmomatic 0.36 (Bolger, Lohse and Usadel, 2014) using the following settings: 1) the Illumina clip function was used to remove adapters, 2) the first six bases were cropped from the start of all paired-end reads, 3) all reads were cropped to 120 bp in length due to lower quality ends (observed using PRINSEQ), and a sliding window was also used to delete bases with a PhredQ score less than 20 with a sliding window of four, and 4) all reads less than 50 bp were discarded, and only paired reads were retained to improve merging of reads during clustering.

Assembly of RADseq data for cycad genera. De novo assembly of the paired-end reads was performed using ipyrad 0.5.13 (Eaton and Overcast, in prep) on a high-performance cluster based at the Royal Botanic Garden Edinburgh using seven nodes, each with 12 cores and 128 GB of RAM, totalling 84 cores and 896 GB of RAM, running for 21 days. In ipyrad all parameters were set to default, except for the following: data type was set to 'pairgbs' (most closely matches ezRAD), bases with a PhredQ score less than 30 were converted to 'N' and reads with 15 or more uncalled bases were discarded. Reads were further filtered for adapter sequences, trimmed, and reads were discarded if they were less than 40 bp in length. The maximum number of uncalled bases in consensus sequences was set to ten for forward and reserve reads. The maximum heterozygotes in consensus sequences was set at eight for both forward and reverse sequences, and the minimum number of samples per locus for output files was set to 4.

Data assembly followed the general ipyrad workflow. Reads were more stringently filtered for presence of adapters (after initial trimming and filtering earlier in Trimmomatic). Next, clusters were identified within samples and consensus base calls were made. Finally, loci were aligned across all of the samples (four species of *Cycas*, and one species each of the nine other cycad genera) and output files were generated, after applying filters as specified in our parameter settings. These settings also included the minimum samples per locus- for example, a generated site is discarded unless it meets the requirement that it is present in a minimum number of samples.

Assembly from population data of *Cycas calcicola*. To further demonstrate the utility of our protocol, we carried out de novo assembly for 72 individuals of *C. calcicola* (one sample failed during sequencing). The minimum number of samples per locus was set to 43 (as opposed to 4 for the genus level assembly, above), so that each site would be present across a minimum of ~ 60% of samples, to reduce missing data.

Mapping of reads to published references. Large cycad genomes (25 - 60 Gbp), present potential problems with overrepresentation of repetitive regions, and for this reason it is important to test the genomic sources and distribution of RADseq reads. To test for overrepresentation reads were mapped against the published reference plastomes and the single mitochondrial genome (Wu et al., 2007 and Wu and Chaw, 2015) (Tables 3 and 4). The reference plastid and mitochondrial genomes were downloaded from NCBI GenBank and the filtered paired end reads were mapped to these references using CLC Genomics Workbench 11.0 (CLC Genomics, 2019; Qiagen Inc., Venlo, the Netherlands) using default parameters: for read alignment mismatch costs = 2, intersection and deletion cost = 3,

length fraction= 0.5, similarity fraction = 0.8 and auto detection of paired distances was allowed.

Phylogenetic analysis of *Cycas calcicola* populations.

The resulting RADseq sequence data provides the first opportunity to investigate the infraspecific relationships between natural populations of *C. calcicola*. Furthermore, this approach can be used to help demonstrate the effectiveness of RADseq in differentiating natural populations. Phylogenetic reconstruction of *C. calcicola* populations was completed using SVDquartet plug-in for PAUP* version 4.0a158 (Swofford, 2002) because of its robust approach in analysing short gene sequences from RADseq data (Liu and Yu 2010, Mirarab et al 2015). Phylogenetic trees were estimated from the concatenated gene sequence alignments using SVDquartets analysis. Settings included exhaustive quartet sampling, 100,000 bootstrap replicates, and the multispecies coalescent tree model. We examined results of all analyses using at least three independent runs for multi-species coalescent analysis by allocating samples to their respective populations. The three separate populations are at Litchfield National Park (including Tolmer Falls sites), Daly River, Katherine CDU, and Spirit Hills.

Results

Number and quality of reads. Sequencing on the Illumina NextSeq 500 platform generated approximately 1.9 to 6.7 million 150 bp paired-end reads per sample (Tables 3, 4 and 5). The number of reads generated varied—with the fewest for *Stangeria eriopus*, and the greatest for *Macrozamia johnsonii.* For *Cycas* (target genus), the number of reads generated showed less variation (1.9 to 2.5 million) and was lowest in *C. taitungensis* and greatest in *C. maconochiei*. The PhredQ Score distribution of the sequencing run measured 75.2% at Q30 or greater, which passed the Illumina sequencing filter. Quality control of reads (measured as PhredQ score in FastQC 0.11.5) indicated that forward reads were of a higher quality with a drop-off after 135 bp, whereas reverse reads were lower quality due to drop-off after 120 bp. Due to this quality drop off, forward and reserve reads were filtered and trimmed to 120 bp. Data accessibility: the data that supported the finding of this study is archived to allow reproducibility of the assembly and filtered sequence reads is accessible from NCBI Sequence Read Achieve, BioSample accession number: PRJNA526348 (Table 2).

Mapping of reads to published references. RADseq reads were mapped against published reference mitochondrial and chloroplast (plastid) genomes. Plastomes ranged in size from 161,815 to 166,431 bp (Table 3). The number of reads mapped to the plastomes varied from 16,292 reads (0.80% of total reads) for *Encephalartos lebomboensis* to *Encephalartos lehmannii* and 221,486 reads (5.82% total number of reads) for *Macrozamia johnsonii* to *M. mountperriensis* (Table 6). The average read depth (Table 3) also varied between the samples and ranged from 10.74 in *E. lebomboensis* to 131.32 in *Cycas armstrongii and* demonstrates that no clusters were over represented. Although the percentage of RADseq reads mapped varied, in all species 89% or greater of the reference was covered and was lowest in *Ceratozamia kuesteriana* (89%) and greatest in *Stangeria eriopus* and *C. armstrongii* (97%).

Reads for *Cycas* spp. were mapped to the mitochondrial genome of *C. taitungensis* which was 414,903 bp (Table 4). The number of reads mapped ranged from 14,672 (0.61% total number of reads) in *C. calcicola* to 26,616 (8.9% total number of reads) in *C. taitungensis*. The number of reads covering the reference mitochondrial genome only varied somewhat between species and was lowest in *C. calcicola* and C. *taitungensis* (62%) and highest in *C. armstrongii* (68%).

De novo assembly of RADseq data. Initial filtering and trimming of the raw Illumina reads were carried out using TRIMMOMATIC. Approximately 65-75% of paired reads were

retained (singletons were removed), each with a minimum PhredQ score of 20 (Table 5). The sample which yielded the lowest number of reads after filtering was *C. taitungensis.* During filtering approximately 1 million reads were discarded for each sample and 3 million reads were removed for *M. johnsonii,* however, *M. johnsonii* remained the taxon with the greatest number of reads overall (Table 5). The number of clusters obtained from de novo assembly ranged from 1.0 to 3.3 million per sample. The number of high-depth clusters (containing six or more reads) ranged from 32,000 in *S. eriopus* to 38,000 in *M. johnsonii* (Table 5). This lower number of high-depth clusters vs initial clusters indicates that there were a high number of clusters with less than six reads, which were discarded due to a higher likelihood of a base being miscalled. The number of recovered loci varied greatly among genera (Table 5), ranging from 1,641 in *C. calcicola* to 1,705 in *C. taitungensis* within *Cycas.* A lower number of loci were recovered for Zamiaceae when compared to Cycadaceae with 125 loci being obtained for *Microcycas calocoma* and 362 for *M.* johnsonii (Table 5).

Example assembly of *Cycas calcicola.* The assembly of 72 samples from natural populations of *C. calcicola* (Table 6), generated 1.7 to 4.7 million reads during sequencing, and most reads passed the ipyrad filter (after trimming). The total number of clusters generated during clustering ranged from 1.3 to 3 million, and the number of high-depth clusters range from 22 to 78 thousand. Overall the assembly generated over three million informative SNPs across the 72 samples, and after final filtering, 4,421 loci were recovered for a minimum of 43 samples per locus (each locus was present for ~60% of samples).

Phylogenetic analysis of *Cycas calcicola.*

The unrooted tree (Figure 1) recovered seven well-supported populations/groups. Spirit Hills, Daly River, Litchfield National Park (NP) and Litchfield Tolmer populations received 100% bootstrap support (BS). Katherine Charles Darwin University site (Katherine CDU) was provided with 99.3% BS and Katherine population and cultivated samples from Katherine TT (Katherine TT CUL) each were provided 90.6% BS. Populations from Katherine and Litchfield national park (NP) were recovered as two separate clades (99.5% and 100%, respectively). Total weight of incompatible quartets was 16.5780 (47.409%), and total weight of compatible quartets was 18.3897 (52.591%).

Discussion

Here we have presented an optimised RADseq protocol used to gain insights into the genetic diversity of cycads. Our results demonstrate that RADseq can successfully be applied across all ten genera of the Cycadales, with sufficient data generated to use this approach for conservation genomics, phylogenetics, and other potential applications.

Assembly of RADseq data. Data was mapped against the reference plastomes and a mitochondrial genome, and showed that less than 8.01% of the total number of reads were mapped. This indicates that neither the plastome or mitochondrial genome were overrepresented in our data, which is further confirmed by the average and maximum read depth (Tables 3 & 4). Additionally, large portions of the reference genomes covered up to 97% of the plastome and 69% of the reference mitochondrial genome. These results are expected with RADseq data as reads will rarely cover the entire reference because of the use of restriction enzymes (Liu and Hansen, 2017). These results indicate that our RADseq protocol is also effective at recovering large portions of the plastome and mitochondrial genome, without reducing the effectiveness and reliability of RADseq for population genetics or phylogenetic inference (Fitz-Gibbon et al., 2017).

De-novo assembly in ipyrad recovered between 125 (*Macrozamia*) to 1,705 (*Cycas*) informative loci, which is the result of several factors: the number of high-depth clusters generated, the number of genetically similar samples included in the assembly and the degree of genetic similarity between species and genera (Table 5). A greater number of

Cycas species were included in the assembly, which are closer genetically (Nagalingum et al., 2011), and is the reason why a greater number of loci were retained for *Cycas,* as with the *Cycas calcicola* example dataset (Table 6). Conversely, fewer loci were recovered for Zamiaceae because of greater genetic distances between genera, and only a single representative species of each genus was included in the assembly. If more samples were included from each genus in Zamiaceae, the resulting number of loci could be greater. Despite the genetic distance among the genera, there was a sufficient number of shared loci recovered between the Zamiaceae and Cycadaceae genera. These results mirror what was found in Myricaceae (Liu et al., 2015) and Diapensiaceae (Hou et al., 2015), as they also found a significant drop in loci recovered in more distantly related taxa, indicating that genetic differences between families would be considerable, as we found between Zamiaceae and Cycadaceae.

The example assembly of *Cycas calcicola* showed a similar result in clustering to that found in the genera dataset by having far fewer high-depth clusters than clusters overall. The assembly generated 4,421 markers across 72 samples using a strict minimum number of samples per locus (to reduce missing data), which required that each locus was present in at least 43 samples (~60%). If the minimum samples per locus was reduced to the default of four, this would further increase the number of loci generated, but also the amount of missing data. This demonstrates that with a good number of samples and a high level of generic similarity, an assembly can generate a good number of loci even with very large genomes. This also appears to have provided sufficient data for coalescent-based analysis since our results were provided with high support (>90% BS) for closely related populations of *C. calcicola*.

Sequencing depth and large genomes. Sequencing resulted in 2.7 to 9.8 million paired-end-reads per sample. Although reads needed to be filtered and trimmed, the sequencing quality was generally high. We aimed to obtain 1 GB per sample to account for the large genome size (25-63 Gbp; Zonneveld, 2012) and overrepresentation of the plastome (Wu and Chaw, 2015). The amount of data (uncompressed) ranged from 1.2 GB for *Stangeria eriopus* to 3.9 GB in *Macrozamia johnsonii*, hence meeting our goal.

One of the main considerations in assembling RADseq data is the clustering of reads for calling consensus sequences and SNPs, as this requires numerous repeat reads to be aligned (Eaton, 2014). In the third step of assembly in ipyrad, if two or more reads aligned, they form a cluster. Subsequently, these clusters are further assessed, and six or more reads (depending on minimum depth clustering depth set) are required for a cluster and its constituent SNPs to be considered reliable—these are termed high-depth clusters (Eaton, 2014). However, in larger genomes, it is less likely that there will be a sufficient number of repeat reads in the sequence data to generate enough high-depth clusters (except for repetitive regions) (Karam et al., 2015). In our study, we found between 1 to 3.3 million clusters in the first clustering step, and 32,000 to 138,000 clusters after selecting only high-depth clusters, indicating that there were many clusters with fewer than six reads. This number of high-depth clusters, while relatively small compared to the initial number, is nonetheless sufficient for downstream phylogenetic and population genetic purposes, especially given that previous work has used significantly fewer markers (Cibrián-Jaramillo et al., 2010, Nagalingum et al., 2011, Meerow et al., 2012, Salas-Leiva et al., 2014, Griffith et al., 2015).

Thus far, RADseq has been utilized in phylogenetics and population genetics for a few plant groups with varying genome sizes (Table 1). The taxa with the smallest genomes (all <1 Gbp) were *Carex* spp., *Sisymbrium austriacum*, *Mimulus* spp, whereas those with the largest genomes include *Diospyros* species (2.40-5.76 Gbp), *Senecio lautus* (4.90 Gbp) and *Pedicularis* species (5.68 Gbp). In our study, RADseq was applied to genomes that are 25 to 63 Gbp - i.e. approximately 4 to 11 times larger than all previous studies. Therefore, we have demonstrated that RADseq can successfully be applied to groups of plants with larger genomes and holds a promise for future applications of RADseq to other plant groups,

especially non-flowering plants with large genomes such as ferns and gymnosperms.

Conclusions. We have demonstrated that RADseq can be applied to organisms with large genomes, such as cycads. This protocol uses high throughput sequencing to recover informative genome-wide markers. RADseq also offers the ability to multiplex and sequence many individuals simultaneously, at relatively low cost. These markers have the potential to be used for population level and for phylogenetic studies, ultimately helping to resolve the relationships among cycads, obtain a better insight into the genetic diversity among the Cycadales species, and to assist in developing informed conservation management plans for cycads and other groups in the future.

Acknowledgements

We wish to acknowledge funding received from the Australian Flora Foundation, the Australasian Systematic Botany Society's Hansjörg Eichler Scientific Research Fund, and The Nature Conservatory - The Thomas Foundation for an Australian Conservation Taxonomy Award. Furthermore, we wish to acknowledge the Biotechnology and Biological Sciences Research Council (BBSRC) UK and the EASTBIO Doctoral Training Partnership for providing a studentship at the University of Edinburgh.

We thank staff at the Royal Botanic Gardens and Domain Trust, in particular, Carolyn Connolly is thanked her support in the molecular laboratory, Hannah McPherson for helping in the initial quality testing of data, and Maureen Phelan and Simon Goodwin for assistance obtaining DNA samples from the living collections. We also thank the various botanic gardens that provided *C. calcicola* tissue. At the Royal Botanic Garden Edinburgh, we wish to thank Laura Forrest and Michelle Hart for their continued support throughout the project. Daren Eaton is acknowledged for assistance with ipyrad: Robert J. Toonen for information about the ezRAD protocol, and Todd McLay for support in quality checking and filtering of RADseq data.

References

Ahrens, C.W., Supple, M.A., Aitken, N.C., Cantrill, D.J., Borevitz, J.O. and James, E.A. (2017). Genomic diversity guides conservation strategies among rare terrestrial orchid species when taxonomy remains uncertain. *Annals of Botany*, 119(8), pp. 1267-1277.

Alexeyenko, A., Nystedt, B., Vezzi, F., Sherwood, E., Ye, R., Knudsen, B., Simonsen, M., Turner, B., de Jong, P., Wu, C.-C. and Lundeberg, J. (2014). Efficient de novo assembly of large and complex genomes by massively parallel sequencing of Fosmid pools. *Bmc Genomics*, 15(1), p. 439.

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G. and Hohenlohe, P.A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews. Genetics*, 17(2), pp. 81-92.

Andrews, K.R., Hohenlohe, P.A., Miller, M.R., Hand, B.K., Seeb, J.E. and Luikart, G. (2014). Trade-offs and utility of alternative RADseq methods: Reply to Puritz et al. Tran. *Molecular Ecology*, 23(24), pp. 5943-5946.

Andrews K. R., Luikart G. (2014). Recent novel approaches for population genomics data analysis. *Molecular Ecology*, 23(7), pp. 1661-1667.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

453 Arnold, B., Corbett-Detig, R.B., Hartl, D. and Bomblies, K. (2013). RADseq underestimates
454 diversity and introduces genealogical biases due to nonrandom haplotype sampling.
455 *Molecular Ecology*, 22(11), pp. 3179-3190.

456 Bayzid, M.S., Hunt, T., Warnow, T. (2014). Disk covering methods improve
457 phylogenomic analyses. *BMC Genomics*, 15(6), p.7.

458 Bolger, A.M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina
459 sequence data. *Bioinformatics*, 30(15), pp. 2114-2120.

460 Calonje, M., Stevenson, D.W. and Stanberg, L. (2013). The World List of Cycas, online
461 edition. Available from http://www.cycadlist.org [Accessed 8 November 2016].

462 Catchen, J.M., Hohenlohe, P.A., Bernatchez, L., Funk, W.C., Andrews, K.R. and Allendorf,
463 F.W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of
464 adaptation in natural populations. *Molecular Ecology Resources*, 17(3), pp. 362-365.

465 Cibrián-Jaramillo, A., Daly, A.C., Brenner, E., Desalle, R. and Marler, T.E. (2010). When
466 North and South don't mix: genetic connectivity of a recently endangered oceanic cycad,
467 *Cycas micronesica*, in Guam using EST-microsatellites. *Molecular Ecology*, 19(12), pp.
468 2364-2379.

469 CLC Genomics Workbench 11.0 (2019). (https://www.qiagenbioinformatics.com/).

470 Chifman, J., Kubatko, L. (2015). Quartet inference from SNP data under the
471 coalescent model. *Bioinformatics*, 30(23), pp. 3317-3324.

472 Davey, J.L. and Blaxter, M.W. (2010). RADSeq: next-generation population genetics.
473 *Briefings in Functional Genomics*, 9(5/6), pp. 416-423.

474 Donaldson, J.S. (2003). *Cycads: status survey and conservation action plan*. IUCN--the
475 World Conservation Union. Available from https://portals.iucn.org/library/node/8203.

476 Eaton, D.A.R. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses.
477 *Bioinformatics*, 30(13), pp. 1844-1849.

478 Eaton, D.A.R. and Overcast, I.A. (2017). *ipyrad: interactive assembly and analysis of*
479 *RADseq data sets*. Available from http://ipyrad.readthedocs.io.

480 Eaton, D.A.R. and Ree, R.H. (2013). Inferring Phylogeny and Introgression using RADseq
481 Data: An Example from Flowering Plants (*Pedicularis*: Orobanchaceae). *Systematic Biology*,
482 62(5), pp. 689-706.

483 Eaton, D.A.R., Hipp, A.L., González-Rodríguez, A. and Cavender-Bares, J. (2015). Historical
484 introgression among the American live oaks and the comparative nature of tests for
485 introgression. *Evolution*, 69(10), pp. 2587-2601.

486 Eaton, D.A.R., Spriggs, E.L., Park, B. and Donoghue, M.J. (2016). Misconceptions on
487 Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants.
488 *Systematic Biology*, 66(3), 399-412.

489 Fitz-Gibbon, S., Hipp, A.L., Pham, K.K., Manos, P.S. and Sork, V.L. (2017). Phylogenomic
490 inferences from reference-mapped and de novo assembled short-read sequence data using
491 RADseq sequencing of California white oaks (*Quercus* section *Quercus*). *Genome*, 60(9),
492 pp. 743-755.

493 Gatesy, J., Springer, M.S. (2014). Phylogenetic analysis at deep timescales: Unreliable gene
494 trees, bypassed hidden support, and the coalescence/concatalescence conundrum.
495 *Molecular Phylogenetics and Evolution*, 80, pp. 231-266.

496 Griffith, M.P., Calonje, M., Meerow, A.W., Tut, F., Kramer, A.T., Hird, A., Magellan, T.M. and
497 Husby, C.E. (2015). Can a Botanic Garden Cycad Collection Capture the Genetic Diversity
498 in a Wild Population? *International Journal of Plant Sciences*, 176(1), pp. 1-10.

499 Guan, R., Zhao, Y., Zhang, H., Fan, G., Liu, X., Zhou, W., Shi, C., Wang, J., Liu, W., Liang,
500 X., Fu, Y., Ma, K., Zhao, L., Zhang, F., Lu, Z., Lee, S.M.-Y., Xu, X., Wang, J., Yang, H., Fu,
501 C., Ge, S. and Chen, W. (2016). Draft genome of the living fossil *Ginkgo biloba*.
502 *GigaScience*, 5(1), p. 49.

503 Guo, Y., Yuan, H., Fang, D., Song, L., Liu, Y., Liu, Y., Wu, L., Yu, J., Li, Z., Xu, X. and
504 Zhang, H. (2014). An improved 2b-RAD approach (I2b-RAD) offering genotyping tested by a
505 rice (*Oryza sativa* L.) F2 population. *Bmc Genomics*, 15(1), pp. 956-956.

506 Guo, W., Grewe, F., Fan, W., Young, G.J., Knoop, V., Palmer, J.D. and Mower, J.P. (2016).
507 Ginkgo and Welwitschia Mitogenomes Reveal Extreme Contrasts in Gymnosperm
508 Mitochondrial Evolution. *Molecular biology and evolution*, 33(6), pp. 1448-1460.

509 Hou, Y., Nowak, M.D., Mirré, V., Bjorå, C.S., Brochmann, C. and Popp, M. (2015).
510 Thousands of RAD-seq Loci Fully Resolve the Phylogeny of the Highly Disjunct Arctic-Alpine
511 Genus *Diapensia* (Diapensiaceae). *PLoS ONE*, 10(10), p. e0140175.

512 IUCN (2016). The IUCN Red List of Threatened Species. [Internet] Available from
513 http://www.iucnredlist.org/about/summary-statistics [Accessed 10 July 2015].

514 Jansen, R.K., Saski, C., Lee, S.-B., Hansen, A.K. and Daniell, H. (2011). Complete plastid
515 genome sequences of three Rosids (Castanea, Prunus, Theobroma): evidence for at least
516 two independent transfers of rpl22 to the nucleus. *Molecular biology and evolution*, 28(1),
517 pp. 835-847.

518 Karam, M.-J., Lefèvre, F., Dagher-Kharrat, M.B., Pinosio, S. and Vendramin, G.G. (2015).
519 Genomic exploration and molecular marker development in a large and complex conifer
520 genome using RADseq and mRNAseq. *Molecular Ecology Resources*, 15(3), pp. 601-612.

521 Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S.,
522 Cooper, A., Markowitz, S. and Duran, C. (2012). Geneious Basic: an integrated and
523 extendable desktop software platform for the organization and analysis of sequence data.
524 *Bioinformatics*, 28(12), pp. 1647-1649.

525 Liu, L., Yu, L., Kubatko, L., Pearl, D.K., Edwards, S.V. (2009). Coalescent methods for
526 estimating phylogenetic trees. *Molecular Phylogenetics and Evolution*, 53(1), pp. 320-328.
527
528 Liu, L., Yu, L., Edwards, S.V. (2010). A maximum pseudo-likelihood approach for
529 estimating species trees under the coalescent model. *BMC Evolutionary Biology*,
530 10(1), pp. 302.
531
532 Liu, L., Jin, X., Chen, N., Li, X., Li, P. and Fu, C. (2015). Phylogeny of *Morella rubra* and Its
533 Relatives (Myricaceae) and Genetic Resources of Chinese Bayberry Using RAD
534 Sequencing. *PLoS ONE*, 10(10), p. e0139840.

535 Liu, S. and Hansen, M.M. (2017). PSMC (pairwise sequentially Markovian coalescent)
536 analysis of RAD (restriction site associated DNA) sequencing data. *Molecular Ecology*

Resources, 17(4), pp. 631-641.

Long-Qian, X., Xue-Jun, G.E., Xun, G., Gang, H.A.O. and Si-Xiang, Z. (2004). ISSR Variation in the Endemic and Endangered Plant *Cycas guizhouensis* (Cycadaceae). *Annals of Botany*, 94(1), pp. 133-138.

Massatti, R., Reznicek, A.A. and Knowles, L.L. (2016). Utilizing RADseq data for phylogenetic analysis of challenging taxonomic groups: A case study in *Carex* sect. Racemosae. *American Journal of Botany*, 103(2), pp. 337-347.

Mastretta-Yanes, A. et al. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, 15(1), pp.28-41.

Meerow, A.W., Francisco-Ortega, J., Colonje, M., Griffith, M.P., Ayala-Silva, T., Stevenson, D.W. and Nakamura, K. (2012). *Zamia* (Cycadales: Zamiaceae) on Puerto Rico: asymmetric genetic differentiation and the hypothesis of multiple introductions. *American Journal of Botany*, 99(11), pp. 1828-1839.

Meyers, B.C., Scalabrin, S. and Morgante, M. (2004). Mapping and sequencing complex genomes: let's get physical! *Nature Reviews. Genetics*, 5(8), pp. 578-588.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17), pp. 541-548.

Mirarab, S., Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12), i44-i52.

Nagalingum, N.S., Marshall, C.R., Quental, T.B., Rai, H.S., Little, D.P. and Mathews, S. (2011). Recent synchronous radiation of a living fossil. *Science*, 334(6057), pp. 796-799.

Narum, S.R., Buerkle, C.A., Davey, J.W., Miller, M.R. and Hohenlohe, P.A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, 22 (11), pp. 2841-2847.

Nazareno, A.G., Dick, C.W. & Lohmann, L.G., (2018). Tangled banks: A landscape genomic evaluation of Wallace's Riverine barrier hypothesis for three Amazon plant species. *Molecular Ecology*, 76(5), p.1-18.

Paun, O., Turner, B., Trucchi, E., Munzinger, J., Chase, M.W. and Samuel, R. (2015). Processes Driving the Adaptive Radiation of a Tropical Tree (*Diospyros*, Ebenaceae) in New Caledonia, a Biodiversity Hotspot. *Systematic Biology*, 65(2), pp. 212-27.

Pellicer, J. et al., (2018). Genome Size Diversity and Its Impact on the Evolution of Land Plants. *Genes*, 9(2), p.88.

Pellicer, J., Fay, M.F. and Leitch, I.J. (2010). The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, 164(1), pp. 10-15.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. and Hoekstra, H.E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE*, 7(5), pp. 1-11.

Raimondo, D.C. and Donaldson, J.S. (2003). Responses of cycads with different life

histories to the impact of plant collecting: simulation models to determine important life history stages and population recovery times. *Biological conservation*, 111(3), pp. 345-358.

Roch, S., Warnow, T. (2015). On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Systematic Biology*, 64(4), pp. 663-676.

Roda, F., Ambrose, L., Walter, G.M., Liu, H.L., Schaul, A., Lowe, A., Pelser, P.B., Prentis, P., Rieseberg, L.H. and Ortiz-Barrientos, D. (2013). Genomic evidence for the parallel evolution of coastal forms in the *Senecio lautus* complex. *Molecular Ecology*, 22(11), pp. 2941-2952.

Roodt, D., Lohaus, R., Sterck, L., Swanepoel, R.L., Van de Peer, Y. and Mizrachi, E. (2017). Evidence for an ancient whole genome duplication in the cycad lineage. *PLoS ONE*, 12(9), p. e0184454.

Ruhsam, M., Rai, H.S., Mathews, S., Ross, T.G., Graham, S.W., Raubeson, L.A., Mei, W., Thomas, P.I., Gardner, M.F., Ennos, R.A. and Hollingsworth, P.M. (2015). Does complete plastid genome sequencing improve species discrimination and phylogenetic resolution in *Araucaria*? *Molecular Ecology Resources*, 15(5), pp. 1067-1078.

Salas-Leiva, D.E., Meerow, A.W., Francisco-Ortega, J., Calonje, M., Griffith, M.P., Stevenson, D.W. and Nakamura, K. (2014). Conserved genetic regions across angiosperms as tools to develop single-copy nuclear markers in gymnosperms: an example using cycads. *Molecular Ecology Resources*, 14(4), pp. 831-845.

Salas-Leiva D.E., Meerow A.W., Calonje M., Griffith M. P., Francisco-Ortega J., Nakamura K., Stevenson D. W., Lewis C. E., Namoff S. (2013). Phylogeny of the cycads based on multiple single-copy nuclear genes: congruence of concatenated parsimony, likelihood and species tree inference methods. *Annals of Botany*, 112(7), pp. 1263-1278.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6), pp. 863-864.

Scott, A.D., Stenz, N.W.M., Ingvarsson, P.K. and Baum, D.A. (2016). Whole genome duplication in coast redwood (*Sequoia sempervirens*) and its implications for explaining the rarity of polyploidy in conifers. *The New Phytologist*, 211(1), pp. 186-193.

Shuguang, J., Yang, Z., Nian, L., Zezheng, G., Qiang, W., Zhenhua, X. and Hal, R. (2006). Genetic variation in the endangered endemic species *Cycas fairylakea* (Cycadaceae) in China and implications for conservation. *Biodiversity & Conservation*, 15(5), pp. 1681-1694.

Sosa, V., Vovides, A.P. and Castillo-Campos, G. (1998). Monitoring endemic plant extinction in Veracruz, Mexico. *Biodiversity and Conservation*, 7(11), pp. 1521-1527.

Swofford DL. (2003). PAUP*: Phylogenetic analysis using parsimony (*and other methods). Version 4 4.0a158.

Toonen, R.J., Puritz, J.B., Forsman, Z.H., Whitney, J.L., Fernandez-Silva, I., Andrews, K.R. and Bird, C.E. (2013). ezRAD: a simplified method for genomic genotyping in non-model organisms. *Peerj*, 1(14), pp. e203-e203.

Tsudzuki, J., Nakashima, K., Tsudzuki, T., Hiratsuka, J., Shibata, M., Wakasugi, T. and Sugiura, M. (1992). Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: nucleotide sequences of trnQ, trnK, psbA, trnI and trnH and the absence of

619 rps16. *Molecular and General Genetics MGG*, 232(2), pp. 206-214.

620 Twyford, A.D. and Friedman, J. (2015). Adaptive divergence in the monkey flower *Mimulus*
621 *guttatus* is maintained by a chromosomal inversion. *Evolution*, 69(6), pp. 1476-1486.

622 Vandepitte, K., Honnay, O., Mergeay, J., Breyne, P., Roldán Ruiz, I. and Meyer, T. (2013).
623 SNP discovery using Paired-End RAD-tag sequencing on pooled genomic DNA of
624 *Sisymbrium austriacum* (Brassicaceae). *Molecular Ecology Resources*, 13(2), pp. 269-275.

625 Yang, Y., Zhou, T., Duan, D., Yang, J., Feng, L. and Zhao, G. (2016). Comparative Analysis
626 of the Complete Chloroplast Genomes of Five *Quercus* Species. *Frontiers in plant science*,
627 07(573), p. 803.

628 Yi, D.-K., Yang, J.C., So, S., Joo, M., Kim, D.-K., Shin, C.H., Lee, Y.-M. and Choi, K. (2015).
629 The complete plastid genome sequence of *Abies koreana* (Pinaceae: Abietoideae).
630 *Mitochondrial DNA Part A*, 5, pp. 2351-2353.

631 Wang, N., Thomson, M., Bodles, W.J.A., Crawford, R.M.M., Hunt, H.V., Featherstone, A.W.,
632 Pellicer, J. and Buggs, R.J.A. (2013). Genome sequence of dwarf birch (*Betula nana*) and
633 cross-species RAD markers. *Molecular Ecology*, 22(11), pp. 3098-3111.

634 Warschefsky, E.J. & Wettberg, von, E.J.B. (2019). Population genomic analysis of mango
635 (*Mangifera indica*) suggests a complex history of domestication. *The New Phytologist*,
636 222(4), 2023-2037.

637 Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L. S., Loopstra, C. A., Vasquez-Gross, H.
638 A., & Holt, C. (2014). Unique features of the loblolly pine (*Pinus taeda* L.) megagenome
639 revealed through sequence annotation. *Genetics*, 196(3), 891-909.

640 Wu, C.-S. and Chaw, S.-M. (2015). Evolutionary Stasis in Cycad Plastomes and the First
641 Case of Plastome GC-biased Gene Conversion. *Genome biology and evolution*, 7(7), pp.
642 2000-2009.

643 Wu, C.-S., Wang, Y.-N., Liu, S.-M. and Chaw, S.-M. (2007). Chloroplast genome (cpDNA) of
644 *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: insights into
645 cpDNA evolution and phylogeny of extant seed plants. *Molecular biology and evolution*, 24
646 (6), pp. 1366-1379.

647 Yang, J., Warnow, T. (2011). Fast and accurate methods for phylogenomic
648 analyses. *BMC Bioinformatics*, 12(9), pp. 4.21.

649 Zgurski, J.M., Rai, H.S., Fai, Q.M., Bogler, D.J., Francisco-Ortega, J. and Graham, S.W.
650 (2008). How well do we understand the overall backbone of cycad phylogeny? New insights
651 from a large, multigene plastid data set. *Molecular Phylogenetics and Evolution*, 47(3), pp.
652 1232-1237.

653 Zhou, T., Zhao, J., Chen, C., Meng, X. and Zhao, G. (2016). Characterization of the
654 complete chloroplast genome sequence of *Primula veris* (Ericales: Primulaceae).
655 *Conservation Genetics*, 8(4), pp. 455-458.

656 Zimin, A., Stevens, K.A., Crepeau, M.W., Holtz-Morris, A., Koriabine, M., Marçais, G., Puiu,
657 D., Roberts, M., Wegrzyn, J.L., de Jong, P.J., Neale, D.B., Salzberg, S.L., Yorke, J.A. and
658 Langley, C.H. (2014). Sequencing and assembly of the 22-gb loblolly pine genome.
659 *Genetics*, 196(3), pp. 875-890.

660 Zonneveld, B.J.M. (2012). Genome sizes for all genera of Cycadales. *Plant Biology*, 14(1),

pp. 253-256.

Zonneveld, B.J.M. (2012). Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. *Nordic Journal of Botany*, 30(4), pp. 490-502.

Zonneveld, B.J.M. and Lindstrom, A.J. (2016). Genome sizes for 71 species of *Zamia* (Cycadales: Zamiaceae) correspond with three different biogeographic regions. *Nordic Journal of Botany*. 34(6), pp. 744-751.

Author Contributions

James A. R. Clugston - writing of paper, performed research and analysed data.

Gregory Kenicer -editing of manuscript and helped analysing data.

Richard Milne - editing of manuscript

Isaac Overcast -constructing, analysing data and writing of papers methodology.

Trevor C. Wilson - analysing data and writing of papers methodology.

Nathalie S. Nagalingum - writing of paper, analysed data and designed research.