TITLE: **Plastome Assembly in the Wax Plants (*Hoya*) *and Near Relatives* (Marsdenieae, Apocynaceae)**

AUTHOR: **Trista Crook**
The Eli Kirk Price Endowed Flora of Pennsylvania Intern

DATE: **April 2017**

ABSTRACT:

The present research sought to assemble the plastomes of 42 species in the Apocynaceae, tribe Marsdenieae. Fast-Plast, a chloroplast assembly pipeline, was used to assemble the plastomes, with alignment to reference plastomes (*Asclepias syriaca* and *Nerium oleander*) and manual editing of the alignments carried out in Geneious. Marsdenieae plastomes proved difficult to assemble, due to presequencing, sequencing, or assembly errors or biological realities. Only 35.7% of the species of Marsdenieae included assembled at all, and no finalized contigs were produced for any species. In most taxa, not all of the contigs aligned with either reference, and nearly all alignments had gaps between contigs. Fast-Plast-assembled contigs and raw reads exhibited many mismatches and gaps when alignments were visually-inspected, the result of errors or biological realities. However, when four metrics—number of contigs assembled, percent of assembled contigs that aligned to each reference plastome, percent coverage of contigs aligned to each reference sequence, and NG50—were considered, most species were deemed to have assembled well, suggesting that these metrics are insufficient to assess assembly quality. Although alignments to the *N. oleander* plastome were slightly better, there was no significant difference between these alignments and those to the *Asclepias syriaca* plastome. When raw reads were aligned to reference plastomes, a pattern of high coverage, accompanied by many gaps and mismatches, suggested that greater genetic distance from a reference sequence may be responsible for at least some of the poor alignment outcomes. Suggestions for improving the alignments and assemblies are given, with an emphasis on possible reasons for the high numbers of variable regions, including assembly or alignment errors, sequencing problems, and taxon-specific chloroplast structure.

Plastome Assembly in the Wax Plants (Hoya) and Near Relatives (Marsdenieae, Apocynaceae)

TABLE OF CONTENTS

INTRODUCTION

Whole genome sequencing data from high throughput sequencing was obtained for 27 species of *Hoya* and closely related genera in the tribe Marsdenieae, Apocynaceae, including *Dischidia* (8 species), *Marsdenia* (3 species), *Oreosparte* (3 species), and *Dischidiopsis* (1 species) with a goal of determining the plastid genome structure of each species. Marsdenieae is pantropical, with most taxa occurring in SE Asia, but is also found in Africa, South America, and Australia (Omlor, 1996; Goyder, 2006). Some authors consider Marsdenieae to be polyphyletic (Potgeiter & Albert, 2001; Senbladd & Bremer, 2002), while others suggest that Marsdenieae is a monophyletic group (Verhoeven, Liede, & Endress, 2003; Meve & Liede, 2004).

This project was undertaken to investigate the following questions:

- What is the gene order and gene content of these Marsdenieae plastomes? How do these plastomes compare to other Apocynaceae plastomes?

- Were there indels, SNPs, and other mutations in Marsdenieae plastomes? If so, where are they located? What do they suggest about the evolutionary history of the group?

Knowledge of the plastome structure in a given taxon can be useful in a wide variety of disciplines from population genetics to evolutionary biology to conservation (Moore et al., 2006; Green, 2007). Gene order, gene content, and gene function can all be understood by analyzing plastomes (Green, 2007; Hall, 2011; Griffiths et al., 2015). Although typically uniparental in inheritance, plastome data can contribute to phylogenetic hypotheses, helping systematists to delimit taxa and understand their evolutionary history.

This research is part of a project that Dr. Tanya Livshultz, my internship co-supervisor, and Dr. Michele Rodda, are collaborating on. Livshultz is an expert in Apocynaceae evolution and classification, with *Dischidia*, *Asclepias*, and *Apocynum* as focus genera, at the Academy of Natural Sciences of Drexel University. Currently based at the Singapore Botanical Garden, Rodda is an expert on *Hoya*.

Structure of the Plastid Genome

Plastid DNA is usually circular (Palmer, 1991). Generally, there is a Large Single Copy (LSC) region followed by an Inverted Repeat (IRb), a Short Single Copy (SSC), and another Inverted Repeat (IRa) (Steele et al., 2012). Single copy refers to the fact that there is only one copy of the DNA in that region, and inverted repeat refers to the fact that these regions are repeated but with the nucleotides in the reverse order of one another. Plastomes range in size from 70-220 kb (Whittall et al., 2010), but most fall between 120 and 160 kb (Palmer, 1991). Size, gene order, and specific genes are largely the same among land plants (Palmer, 1991; Steele et al., 2012). The plastome of *Asclepias syriaca* is 158,719 bp (GenBank KF 386166.1), whilst that of *Nerium oleander* is 154,903 bp (GenBank KJ953907.1). Some genes and regions of the chloroplast genome evolve much faster, and are thus more variable, than other genes and regions (Palmer, 1991; Moore et al., 2007; Nock et al., 2011). Genes with essential functions, such as photosynthesis, tend to be more conserved than other genes (Palmer, 1991; Kim & Lee, 2004). Genes in the IR regions also tend to evolve more slowly than those in the LSC and SSC (Kim & Lee, 2004). IR regions commonly expand and contract as indels become part of the genome (Palmer, 1991).

DNA Sequencing and Plastome Assembly

Sequencing is a process in which the order of the nucleotides in a DNA molecule is determined by performing a series of chemical reactions. In Next Generation sequencing (NGS, massive parallel sequencing, or high throughput sequencing) millions of copies of small fragments of DNA called reads, which range from 30-400 bp in length, are produced (Simon et al., 2009). The process is relatively affordable but can be prone to errors and makes plastome assembly difficult.

Reads are assembled into one or more contigs (contiguous DNA segments) by a suite of software programs contained within an assembly pipeline. Another software program is used to align the contigs to a reference sequence, i.e. an already completely sequenced plastome of a closely related species, in order to see if the bases are correct or not and to search for possible mutations within the DNA sequence.

Assessing Assembly Quality

There are many different ways of evaluating the quality of an assembly. The most robust of these analyze many aspects of the assembly at once and require knowledge of at least one programming language. For example, BUSCO compares the assembly to a set of typically well-conserved genes within a group of organisms and searches the assembly for those genes (Simão et al., 2015). However, simpler metrics are often used, such as NG50, the number of contigs assembled, the length of assembled contigs, coverage of contigs over the reference sequence(s), and the number of ambiguous bases (Salzberg et al., 2012; Wysocki et al., 2014). NG50 is the median length of the contigs at which half of the estimated plastome size has been assembled (Earl et al., 2011).

METHODS

The following Marsdenieae taxa assemblies were analyzed for this project: *Dischidia acutifolia*, *Dischidia hirsuta*, *Dischidia major*, *Dischidia milnei*, *Dischidia nummularia*, *Dischidiopsis parasitica*, *Hoya bakoensis*, *Hoya diversifolia*, *Hoya exilis*, and *Oreosparte celebica*. *Alyxia siamensis* and *Aspidosperma cruentum* (Alyxieae and Aspidospermae) were added as control taxa as they had been previously assembled successfully (Straub, n.d.).

Reads were assembled by the assembly pipeline Fast-Plast, run on the University of Drexel cluster, Proteus, using command line entry. Fast-Plast contains the following programs: Trimmomatic, Bowtie, SPAdes, Afin, and Plastome Finisher. Fastq files, which contain the raw reads, were added to the Proteus cluster. Trimmomatic removes adapters from the raw reads, but this process was skipped since my data had already been trimmed when I received them. Bowtie 2 aligns the reads using 320 angiosperm plastome reference sequences from GenBank and Verdant (Langmead & Salzberg, 2012; McKain & Wilson, 2016). Additional reference sequences can be added. In the process, it removes mitochondrial and nuclear DNA. Next, the aligned reads pass through SPAdes, a program that joins overlapping reads together to form contigs. In order to be assembled into contigs, reads must be *k* base pairs long, with k set by the user. SPAdes stores contigs in fasta format, which contains the sequences, lengths, and average coverage for each contig. SPAdes is unable to make complete contigs when contigs do not overlap enough or are too dissimilar. Afin removes the ends of the contigs, and starts extending them by adding trimmed reads and checking to see if they overlap at a minimum threshold until all of the SPAdes contigs are assembled into fewer (ideally one) contig(s). This process is repeated until the best possible contigs are created, and is necessary because the ends sometimes have errors that SPAdes cannot detect. Finally, Plastome Finisher compares the contigs to reference sequences, annotating the plastome with the gene order and producing contigs for the different regions of the plastome (J. Teisher, personal communication, August 26, 2016).

Geneious version 5.6 (http://www.geneious.com, Kearse et al., 2012) was used to align the contigs to a reference plastome for adjusting the assembly parameters by hand as necessary. The following default parameters were used: Medium Sensitivity/Fast, Maximum (slowest) Fine Tuning. Under these parameters, the following advanced criteria are automatically set by Geneious: Allow gaps with a maximum of 15% per read and size of 50 bp, a word length with minimum overlap of 25 (with words repeated more than 10 times ignored), a maximum of 15%

mismatches per read, paired read distances used to improve assembly, a maximum gap size of 50 bp, a minimum overlap identity of 80%, and index word length of 12, and a maximum ambiguity of 4. *A. syriaca* and *N. oleander* (Apocynaceae) plastomes were downloaded from GenBank, https://www.ncbi.nlm.nih.gov/genbank/, (KF 386166.1 and KJ953907.1, respectively) as these species have been used as references in several other studies of Apocynaceae taxa (Straub et al., 2012, Straub et al., 2013; Straub et al., 2014).

Masked areas in a consensus sequence were filled by the letter N, which indicates that the correct sequence order is unknown at the given location, and helps improve the alignment. Gaps resulting from lack of coverage or a possible insertion were masked, while gaps due to a possible deletion were removed. However, when many of the raw reads did not have a base in the position of a gap in the consensus sequence, the gap was kept because an insertion might have occurred. Reads with problematic end sequences and a lack of coverage were either deleted or the gap was removed from the consensus sequence.

Areas of the consensus that had many mismatched base pairs were closely inspected. For example, if the majority of reads indicated that the base was an A, then the consensus sequence would be edited to an A at that position. However, if it was difficult to determine the most common base, then an N would be used instead. When half of the raw reads indicated that a certain base should be at a given position and the other half indicated another base should be there, Geneious used an R to indicate that the base could be A or G or a K to indicate that a base could be C or T. The base at that position in the reference was used in the consensus in such cases.

When mismatches occur at the ends of reads or contigs, it is more likely that they are due to a sequencing or assembly error rather than a biological difference between the reference and query sequences. In this study if there were more reads that aligned with the reference than ones that did not, then the ends were likely a sequencing error, but if the opposite were true, then the mismatched end could be an insertion, requiring closer inspection (J. Teisher, personal communication, November 30, 2016).

Three metrics that were used to evaluate the quality of assemblies (NG50, percentage of contigs that aligned to each reference sequence, and percent coverage of contigs over each reference sequence) were analyzed in R (R Core Team 2017) using R commander (Fox & Bouchet-Valat, 2005). Data were checked for normality and homogeneity of variances prior to further analyses. If these conditions were met, data was subjected to independent samples t-tests. Otherwise, Mann-Whitney U tests were performed.

To determine if the problems observed in the contig alignments were due to errors or an unusual chloroplast structure in these species, the raw reads from previously-assembled Apocynaceae species (Straub et al., 2014) were used to assemble contigs using Fast-Plast. These contigs were then aligned to the *N. oleander* reference plastome and evaluated in Geneious.

RESULTS

Alyxia siamensis, *H. bakoensis*, *D. milnei*, *O. celebica*, and *Aspidosperma cruentum* contigs did not align with the *Asclepias syriaca* ndhF gene. No contigs from *D. hirsuta* aligned with the *N. oleander* ndhF gene. No *Alyxia siamensis* and *Aspidosperma cruentum* contigs aligned with the *Asclepias syriaca* plastome.

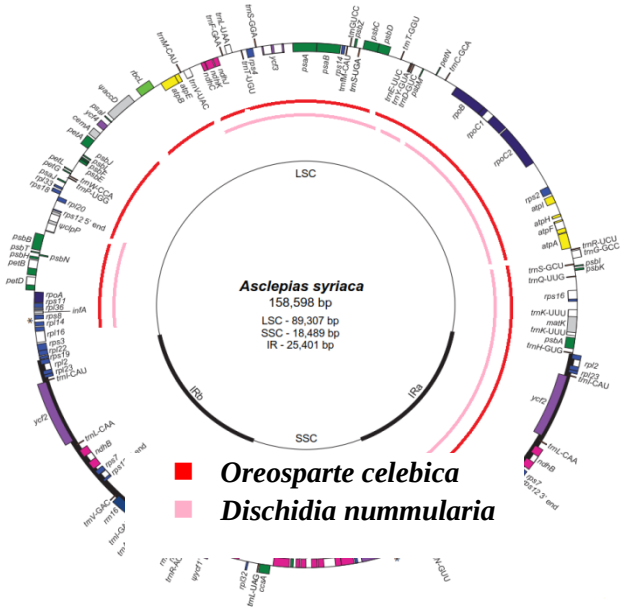
Only a few of the Afin-assembled contigs for nearly all taxa aligned with either reference plastome, and all but one species (*Aspidosperma cruentum*) assembled into more than one contig in Afin (Fig. 2). Generally, alignments with reference sequences resulted in consensus sequences with many gaps and mismatches, regardless of whether contigs or raw reads were used. Most of these gaps and mismatches were in intergenic regions. *D. hirsuta* contigs did not align with a partial sequence for the *D. hirsuta* matK gene downloaded from GenBank (HQ 327590.1). Among the *N. oleander* plastome alignments, *D. hirsuta* had many mismatches and the largest gaps between contigs seen in any species, with approximately 65,280 bp between two of the contigs and a lack of coverage for the first 32,410 bp. Among all alignments with all reference sequences, *D. milnei* did not align with the *N. oleander* plastome and had the largest gaps between contigs among the *Asclepias syriaca* alignments, with the alignment not beginning until 130,719 bp in.

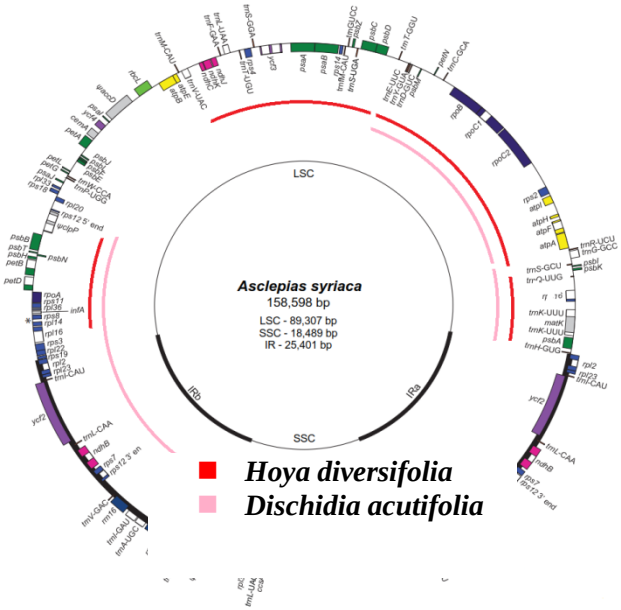
When raw reads were aligned to the *Asclepias syriaca* plastome, Marsdenieae species had mean depth values ranging from 135.7X to 1051.1X, whilst *Alyxia siamensis* and *Aspidosperma cruentum* had mean depths of 35X and 31.6X, respectively. Raw reads aligned to the *N. oleander* plastome for Marsdenieae species had mean depths from 141.5X to 1089.6X, whilst *Alyxia siamensis* and *Aspidosperma cruentum* had mean depths of 42.1X and 37.7X, respectively. Despite these great depths, gaps were still common in all alignments.

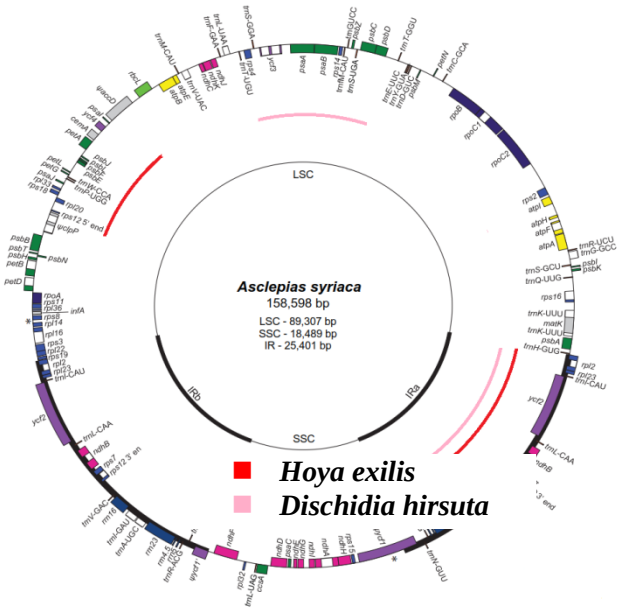
There was at least some coverage of each region of the *Asclepias syriaca* plastome by Marsdenieae contigs (Fig. 1). (Note that this coverage observes only gaps between contigs and not gaps within contigs.) The IRa was the best covered region among all species except *H. diversifolia*, which exhibited no coverage for the region. The IRb was either not covered or barely covered by most species' contigs, and ndhF and ycf1 were not covered by five species (62.5% of species that aligned with the *Asclepias syriaca* plastome). Only one species, *D. acutifolia*, aligned over the entire IRb. Contigs from all species covered at least part of the IRa, except for *D. acutifolia*. There was no observable pattern of coverage for contigs assembling over the LSC or the SSC other than most species had at least some coverage in the two regions. Contigs from nine species, including two from Straub (n.d.), aligned with the *N. oleander* plastome. Only three of these species (*H. bakoensis*, *Aspidosperma cruentum*, and *D. hirsuta*) had coverage in the IR regions and in ycf1, and only one (*Aspidosperma cruentum*) had coverage at ndhF. Coverage in the LSC was spotty, though the contigs of most Marsdenieae species covered some portion of the region. There were no significant differences between *Asclepias syriaca* and *N. oleander* alignments on any measure (NG50, coverage of plastome, percent of contigs aligned), with and without *Alyxia siamensis* and *Aspidosperma cruentum* included (See Appendix A).

Table 1. Overall assembly quality for each species based on four metrics. Species that assembled well met three of the following criteria: (1) an NG50 ≥ 50% of both reference plastome lengths, (2) percent coverage ≥ 60% for both references, (3) number of contigs <21 (i.e. ½ the largest number of contigs), and (4) ≥ 60% of contigs aligned to both references.

Assembled Well	**Assembled Poorly**
Aspidosperma cruentum	*Alyxia siamensis*
Dischidia nummularia	*Dischidia acutifolia*
Hoya bakoensis	*Dischidia hirsuta*
Hoya diversifolia	*Dischidia major*
Oreosparte celebica	*Dischidia milnei*
	Dischidiopsis parasitica
	Hoya exilis


Asclepias syriaca
158,598 bp
LSC - 89,307 bp
SSC - 18,489 bp
IR - 25,401 bp
LSC
IRb
IRa
SSC
Oreosparte celebica
Dischidia nummularia


Asclepias syriaca
158,598 bp
LSC - 89,307 bp
SSC - 18,489 bp
IR - 25,401 bp
LSC
IRb
IRa
SSC
Hoya diversifolia
Dischidia acutifolia


Asclepias syriaca
158,598 bp
LSC - 89,307 bp
SSC - 18,489 bp
IR - 25,401 bp
LSC
IRb
IRa
SSC
Hoya exilis
Dischidia hirsuta

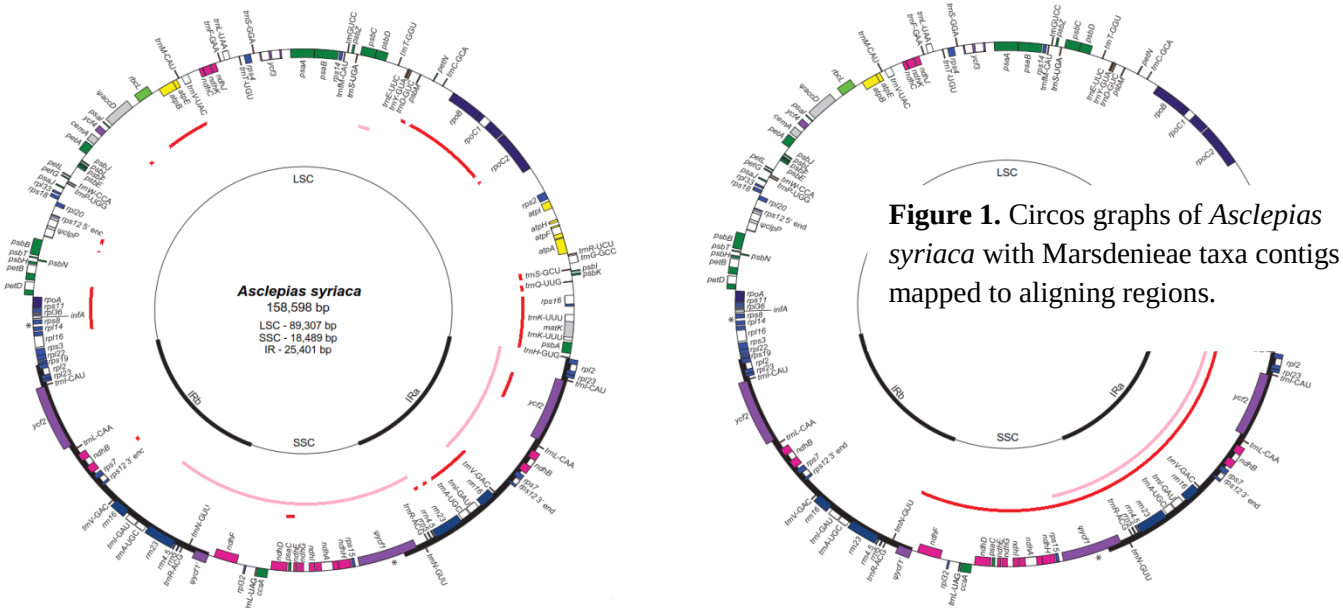
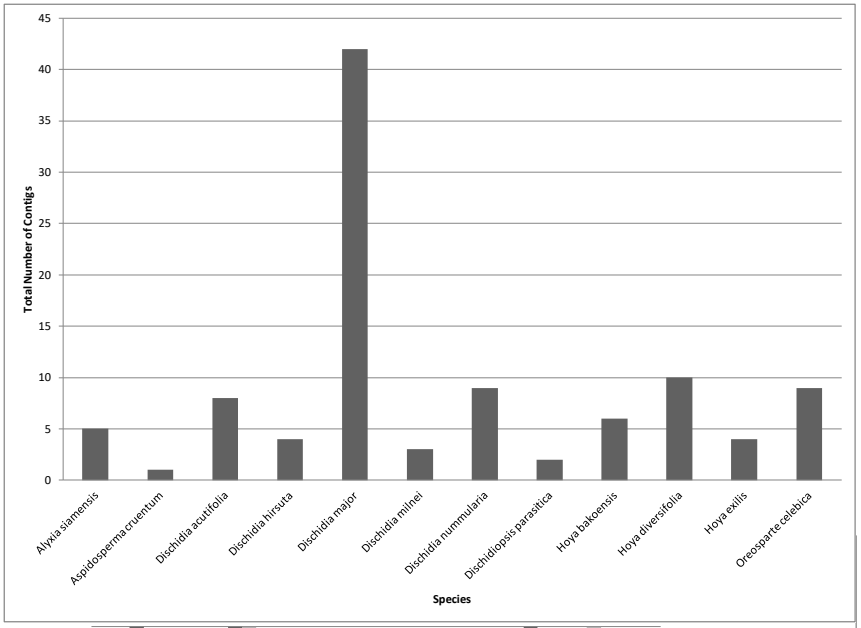
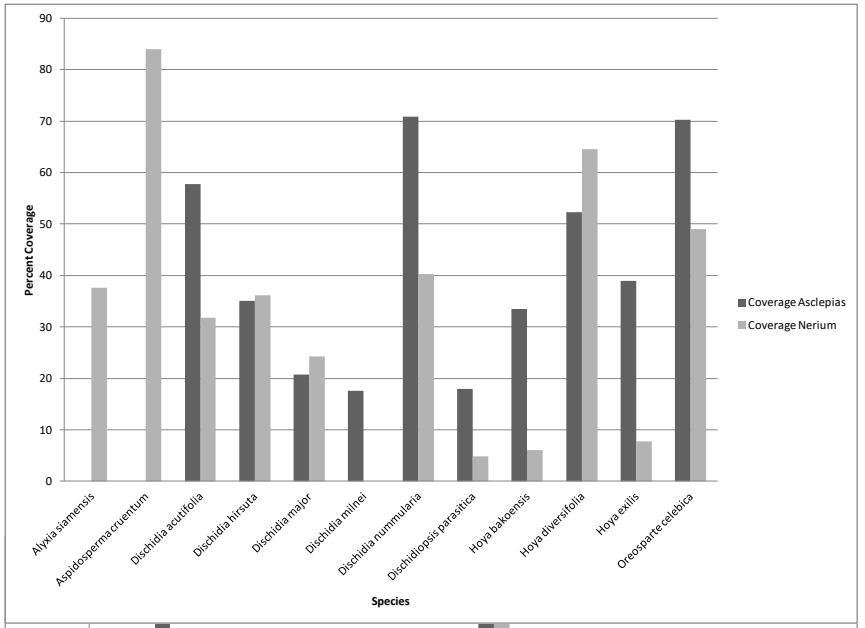


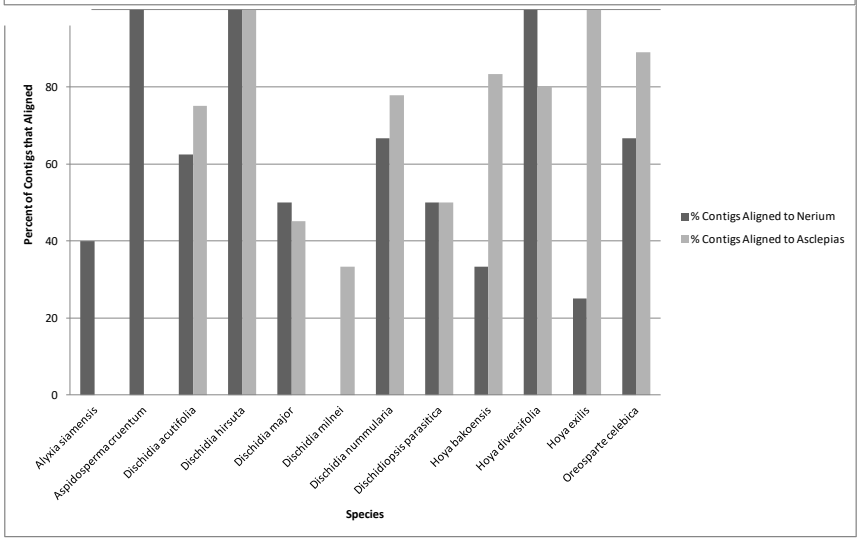
Figure 1. Circos graphs of *Asclepias syriaca* with Marsdenieae taxa contigs mapped to aligning regions.

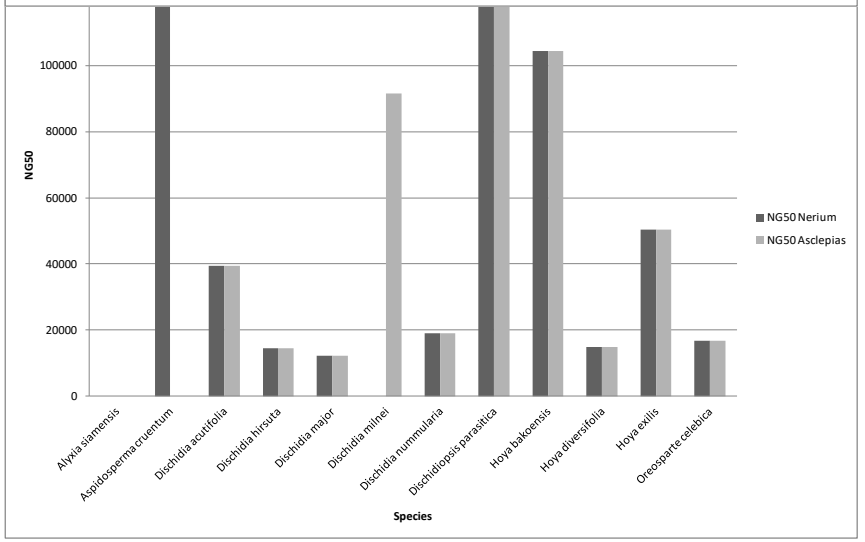
■ ***Dischidia major***
■ ***Hoya bakoensis***
■ ***Dischidiopsis parasitica***
■ ***Dischidia milnei***

(b)









the
of contigs
each
for each

DISCUSSION

Sequencing Errors

Sequencing errors occur during the process of sequencing DNA, and can result from issues with methodology, chemical processes, DNA damage from the sequencing process itself, and biological realities (Dohm et al., 2008; Costello et al., 2013; Schirmer et al., 2015). The short reads generated from NextGen sequencing may make it impossible to correctly assemble certain regions, e.g., Steele et al. (2012) were unable to assemble indels greater than 50 bp from 80-120 bp reads. The short reads in this study (100 bp long) may have contributed to assembly difficulty, particularly as indels longer than 100 bp are impossible for Geneious to correctly align to a reference (Kearse, 2015).

Incorrect base calls are more common than bases being inserted or deleted from sequences generated by Illumina sequencing technology (Kelley, Schatz, & Salzberg, 2010; Minoche, Dohm, & Himmelbauer, 2011; Schirmer et al., 2015), which is likely the method used to sequence the Marsdenieae taxa, according to Geneious predictions. (The sequencing platform that was used has not been provided by the company that performed the sequencing.) The likelihood that a base is called correctly is determined by the sequencing software and reported as confidence means. Marsdenieae data confidence means ranged from 35.4 to 36.4, indicating that the probability of the base being correctly called was 99.9% (Dohm et al, 2008; Minoche, Dohm, & Himmelbauer, 2011). This suggests that sequencing errors were minor, not affecting the sequence in a significant way, as is generally the case (Médigue et al., 1999). However, Dohm et al. (2008) found that Solexa software is highly inaccurate in determining these values.

Assembly and Alignment Difficulties and Errors

Strong evidence of a biological reality or assembly error is based on the fact that the last piece of software in the Fast-Plast assembly pipeline, Plastome Finisher, would not run. This program could not determine the location of the chloroplast genome regions or the gene order of the Afin-assembled contigs, which could be the result of a recent update to Afin or an unusual chloroplast structure in Marsdenieae (J. Teisher, personal communication, August 26, 2016). Wysocki et al. (2014) found that among different assemblers, SPAdes had the greatest amount of ambiguous and missing bases, despite producing the longest contigs, a further indication that many metrics must be taken into account when assessing assembly quality and suggesting that SPAdes may be at least partially responsible for problems in the assembly of Marsdenieae taxa.

Problems with the Ends of Reads and Contigs

The ends of reads are often riddled with errors as a result of sequencing methods (Dohm et al., 2008; Kelley, Schatz, & Salzburg 2010; Schirmer et al., 2015), which could carry over to contigs. However, biological realities and assembly problems can also lead to contig ends with many errors, mismatches, or indels (J. Teisher, personal communication, December 21, 2016). For instance, ends of contigs or reads can contain codons for adjacent genes, resulting in inaccurate alignments when the reference sequence has a different gene order (Hall, 2011; Straub, 2012). Marsdenieae contig ends frequently had many mismatches, which, considering the high confidence means are most likely not due to sequencing errors. Many of these mismatches were located at roughly the same position in multiple species, suggesting that these regions contain actual indels or SNPs. However, it is possible that a systematic error in the

sequencing, assembly, or alignment process led to the same error being repeated multiple times in the same location of the plastome.

Evaluating Assembly Quality

When evaluated using NG50, percent coverage of reference plastome, number of contigs that assembled, and percent of contigs that aligned with each reference plastome, only 42% of species assembled well. Quality of assemblies varied among species when different metrics were considered individually (Fig. 2). Visual observations of the alignments suggested that the many gaps and mismatches observed were indicative of poor alignments for all species. Thus, the metrics chosen to evaluate assembly quality may be overly simplistic and inadequate to evaluate overall assembly quality.

Although *Aspidosperma cruentum* and *Dischidiopsis parasitica* had few contigs and high NG50 values (Fig. 2), these values are not necessarily indicative of good assemblies, as regions distant from one another might have been assembled (Salzberg et al., 2012). Evidence of this occurring in *Aspidosperma cruentum* stems from the fact that this species did not align with the *Asclepias syriaca* plastome. In *Dischidiopsis parasitica*, only a small portion of one of the long contigs assembled actually aligned with the reference plastomes. *D. hirsuta* and *D. milnei* had the largest gaps between contigs when aligned to the *N. oleander* plastome and *Asclepias syriaca* plastome, respectively (Fig. 2). Additionally, *D. hirsuta* contigs would not align with any genes, including a partial matK gene from another *D. hirsuta* individual in GenBank (HQ 327590.1). These issues suggest DNA preparation errors (e.g. contamination), sequencing errors, assembly errors, or an especially highly rearranged plastome (Steele et al., 2012).

Coverage of the *Asclepias syriaca* plastome was spotty for *D. major* (Fig. 1). Larger numbers of contigs are associated with increased assembly difficulty and errors (Steele et al., 2012). By this criterion, *D. major* is likely to have among the most inaccurate assemblies since Afin produced 42 contigs for the species' plastome. That neither of the control taxa (*Alyxia siamensis* and *Aspidosperma cruentum*) aligned with the *Asclepias syriaca* plastome suggests that the errors seen in all taxa are largely based on difficulties within the assembly pipeline. This still does not dismiss the possibility of a highly rearranged plastome in the Marsdenieae, as gaps and mismatches in these alignments could still result from biological realities. However, visual inspection of alignments with the *N. oleander* plastome appearing to have fewer gaps and mismatches are more in line with presequencing or sequencing errors and an unusual plastome structure in the Marsdenieae.

Effects of Biological Realities on Plastome Assembly

Pseudogenes, highly variable regions of the plastome, inversions, inverted repeats, indels, and SNPs make assembly difficult, resulting in errors or making it difficult to distinguish errors from authentic biological occurrences. Areas that cannot be assembled are often those that are evolving the fastest and thus are highly variable among taxa (Straub et al., 2012). In this study, most gaps and mismatches were in intergenic regions, a finding that is in line with several studies (Nock et al., 2011; Parks, Liston, & Cronn, 2010). This is not surprising considering that intergenic regions tend to be highly variable (Palmer, 1991), and highly variable regions have many mismatches and ambiguities (J. Teisher, personal communication, November 30, 2016). However, Parks, Liston, & Cronn (2010) did not have trouble assembling intergenic spacers.

Repeats and inversions are particularly difficult to assemble, the former leading to gaps in reads (Earl et al., 2011; Nakamura et al., 2011; Salzberg et al., 2012). Furthermore, repeats are common in boundary genes (Kim & Lee, 2004; Yue et al., 2008; Straub et al., 2014) and at inversions, leading to assembly errors and difficulties (Palmer, 1991; Rabinowicz & Bennetzen, 2006). This can explain the lack of coverage in IR regions across the reference sequences, particularly as Fast-Plast does not work well when there are inversions at the IR boundaries (J. Teisher, personal communication, October 5, 2016). Gaps in reads due to repeats could account for assembly errors as well, and explain the high number of gaps and mismatches in the alignments.

The genes ycf1 and ndhF are notoriously difficult to assemble, as they move between the IRb and SSC, sometimes straddling both regions (Davis & Soreng, 2010, J. Teisher, personal communication, August 26, 2016) or missing altogether (Steele et al., 2012). Kim & Lee (2004) observed that ndhF and ycf1 overlap on the border of the IRb and the SSC in *Arabidopsis*, a finding that supports the lack of any alignment of Marsdenieae contigs to the ndhF region of the *N. oleander* complete plastome. Furthermore, the length of IR regions varies among taxa, resulting in pseudogenes at IR and SSC boundaries (Kim & Lee, 2004). In this study, ndhF and ycf1in the target taxa did not align with the *Asclepias syriaca* plastome in 62.5% of the species that aligned, suggesting that these genes have shifted to different locations in at least some of the Marsdenieae plastomes, there are inversions at the boundary of the IRb and SSC, and/or they might be pseudogenes.

The *Asclepias syriaca* plastome is notoriously difficult to assemble. Even after Sanger sequencing *Asclepias syriaca*, Straub et al. (2011) were unable to assemble certain portions of its plastome, including the purported pseudogene ycf1. They concluded that repeats and interactions between pseudogenes found in the plastid and mitochondrial DNA make assembly from short reads impossible (Straub et al., 2011). It is possible that Marsdenieae has a similarly complex plastome that is difficult to assemble. *N. oleander* may prove a better reference sequence (T. Livshultz, personal communication, September 21, 2016), as supported in the present work by the slightly improved, though non-significant, alignments seen with *N. oleander*, although *Asclepias syriaca* is in the same subfamily as Marsdenieae (Asclepioideae), whilst *N. oleander* is in the Apocynoideae, an older, more distantly related subfamily. Inversions might be common in the Marsdenieae as they are in other Apocynaceae plastomes (T. Livshultz, personal communication, October 5, 2016). Marsdenieae taxa may have pseudogenes as well, which cannot be misassembled when coverage is too high, especially when paired end reads are used (Li & Homer, 2010; Straub et al., 2012), possibly explaining the poor assemblies in this study despite the high coverage.

Divergence from the reference can make assembly more difficult, error-prone, and result in an incomplete plastome (Steele et al., 2012; Straub et al., 2012). Nock et al. (2011) found more gaps in the consensus sequence when target references were aligned to a more distantly related species, even though coverage exceeded 100X. The high coverage accompanied by many gaps in this study could indicate that this tribe has highly variable plastomes that are so dissimilar from *Asclepias syriaca* and *N. oleander* that they do not align well (or at all) across the entire sequence. Furthermore, indels and repeats are hard to assemble and distinguish using a reference

(Nock et al., 2011; Straub et al., 2011). Yet, several studies suggest that, in many cases, the relatedness of the reference to the target species does not significantly affect assembly success (Parks, Liston, & Cronn, 2010; Steele et al., 2012; Straub et al., 2012).

Future Work

Resequencing using Roche 454 or Pacific BioSciences RS II or another NGS platform that produces longer or more accurate reads followed by Sanger sequencing of difficult to assemble regions could help improve the assembly. *De novo* assembly of the reads for each Marsdenieae taxon in the study could also improve outcomes and clarify if the large number of indels were real, and potentially provide evidence for the possibility of the Marsdenieae plastome being largely reorganized in comparison to the reference plastomes used in this study (Straub et al., 2012). Another solution would be to fill in missing target sequence in the consensus sequence with the bases from the reference sequence, and then align *de novo* contigs to this newly generated "chimeric reference" (J. Teisher, personal communication, February 17, 2017; Whittall et al., 2010; Parks, Liston, & Cronn, 2010; Straub et al., 2013). Straub et al. (2012) suggest sequencing one species to a greater depth than others in a taxon, followed by assembly of the plastome using increasingly dissimilar references and *de novo* assembly. The plastome sequence for that species would then be used as the reference for the others in the group.

Better alignments can be achieved by adjusting the parameters used during assembly and alignment (J. Teisher, personal communication, September 7, 2016). Smaller deletions can be aligned if the maximum gap length is increased, and indels, duplications, and inversions can be detected when other alignment tools within Geneious are used (Kearse, 2015). Using less stringent criteria, such as those used by Straub et al. (2013), might help target sequences align with reference sequences when they originally did not, while stricter criteria could produce fewer gaps and mismatches. Since it seems likely that the sequence divergence from the target negatively affected assemblies in this study, looser parameters may result in an improved final assembly.

CONCLUSION

The original research questions regarding gene content, gene order, and evolutionary history in Marsdenieae were unable to be answered due to difficulties with plastome assembly. This research indicates that more work is required to determine the plastid structure of these Marsdenieae species. By visual inspection of alignments abundant mismatches, gaps, and missing sequence suggested poor assemblies. Different measures of assembly quality indicated that different species had better assemblies than others for one metric, but not the same species when all metrics were considered. The lack of more robust metrics with more consistent results makes it difficult to determine which assemblies were better than others overall. Only a few species aligned well with reference sequences based on the NG50, number of contigs assembled, percent of contigs that aligned with reference sequences, and percent coverage of the reference sequences by aligned contigs. All species aligned poorly based on qualitative visual assessments of alignments with many mismatches and gaps. The assembly pipeline failed to assemble contigs for most of the species, with no finalized contigs produced in the last stage of the assembly for any species. Multiple gaps and mismatches were observed in the alignments to reference sequences, with alignments to *N. oleander* proving to have slightly fewer poorly

aligned regions, though not significantly so, than those to *Ascelpias syriaca*, likely due to a less rearranged plastome since *N. oleander* is more distantly related to Marsdenieae than *Asclepias syriaca*. Analyzing gaps between contigs revealed that all regions of the plastome posed assembly difficulties.

Sanger sequencing, adjustment of the assembly software, use of more closely-related and less rearranged plastome references, and *de novo* assembly can help solve these problems. For particularly difficult to assemble and align species, such as *D. hirsuta*, *D. major*, and *D. milnei*, high throughput sequencing might need to be repeated to assure that there was no contamination or other sequencing issues.

Assembly, DNA extraction and processing, sequencing, or biological realities could explain assembly difficulties. Assembly errors due to a highly rearranged plastome in the Marsdenieae may be responsible for most errors, considering that Apocynaceae plastomes are known to be difficult to assemble due to an unusual plastid genome structure. Highly divergent reference sequences may also account for many of the errors, as *N. oleander* is not closely related to the Marsdenieae and *Asclepias syriaca* has a highly rearranged plastome. Sequencing seems least likely to explain errors as confidence means were high. It was not possible to ascertain if there were errors in DNA preparation methods as such errors would be masked by others and only obtainable through experimental analysis.

GLOSSARY

DNA: building block of life, composed for four nucleotides bases, adenine (A), guanine (G), cytosine (C), and thyamine (T), with A & T bonded together and C & G bonded together, and a phosphate backbone.
Indel: collective term for insertion or deletion of nucleotide bases relative to another sequence.
Inversion: occurs when a double-stranded DNA sequence is flipped 180^{o} and reinserted into the genome.
Pseudogene: regions of DNA with a similar sequence to a gene but with mutations that prevent them from being functional.
Repeat: repetition of one or more nucleotide bases in succession.
Single Nucleotide Polymorphism (SNP): a nucleotide at a given position differs from that in another sequence.
(Graur & Li, 2000; Griffiths et al., 2015)

ACKNOWLDGEMENTS

Thank you to Jordan Teisher at the Academy of Natural Sciences for his mentoring, expertise, and assistance with plastome assembly and methodology. I would also like to thank Tanya Livshultz at the Academy of Natural Sciences for suggesting the project and providing advice on how to improve assemblies, Michele Rodda at Singapore Botanic Gardens for providing the raw data. Finally, I would like to thank Cindy Skema at the Morris Arboretum for her assistance, advice, and expertise in writing and revising this report.

REFERENCES

Costello, M., Pugh, T. J., Fennell, T. J., Stewart, C., Lichtenstein, L., Meldrim, J. C., . . . Getz, G. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research, 41*(6). doi:10.1093/nar/gks1443

Davis, J. I., & Soreng, R. J. (2010). Migration of endpoints of two genes relative to boundaries between regions of the plastid genome in the grass family (Poaceae). *American Journal of Botany, 97*(5), 874-892. doi:10.3732/ajb.0900228

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research, 36*(16). doi:10.1093/nar/gkn425

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D. W., Fass, J., . . . Paten, B. (2011). Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. *Genome Research, 21*(12), 2224-2241. doi:10.1101/gr.126599.111

Fox, J., & Bouchet-Valat, M. (2017). Rcmdr: R Commander. R package version 2.3-2.

Goyder, D. J. (2006). An overview of Asclepiad biogeography. *Taxonomy and Ecology of African Plants, Their Conservation and Sustainable Use.* Royal Botanic Gardens, Kew, 205-214.

Green, P. (2007). 2x genomes - Does depth matter? *Genome Research, 17*(11), 1547-1549. doi:10.1101/gr.7050807

Graur, D. & Li, W. (2000). *Fundamentals of Molecular Evolution* (Second ed.). Sunderland, MA: Sinauer Associates, Inc. Publishers.

Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., & Doebley, J. (2015). *Introduction to Genetic Analysis* (Eleventh ed.). New York, NY: W. H. Freeman and Company.

Hall, B. G. (2011). *Phylogenetic Trees Made Easy: A How-To Manual* (Fourth ed.). Sunderland, MA: Sinauer Associates, Inc. Publishers.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjies, P., & Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics, 28*(12), 1647-1649.

Kearse, M. (2015, July 28). Message posted to: https://support.geneious.com/hc/en-us/articles/227535048-What-s-the-difference-between-Pairwise-Multiple-alignment-de-novo-Assembly-and-Map-to-Reference-

Kelley, D. R., Schatz, M. C., & Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome Biology, 11*(11). doi:10.1186/gb-2010-11-11-r116

Kim, K. J., & Lee, H. L. (2004). Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Research, 11*(4), 247-261. doi:10.1093/dnares/11.4.247

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357-U354. doi:10.1038/nmeth.1923

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics, 11*(5), 473-483. doi:10.1093/bib/bbq015

McKain, M. R., & Wislon, M. (2016). Fast-Plast: Rapid *de novo* assembly and finishing for whole chloroplast genomes. Retrieved from *GitHub Repository*, https://github.com/mrmckain/Fast-Plast

Médigue, C., Rose, M., Viari, A., & Danchin, A. (1999). Detecting and analyzing DNA sequencing errors: toward a higher quality of the *Bacillus subtilis* genome sequence. *Genome Research, 9*(11), 1116-1127. doi:10.1101/gr.9.11.1116

Meve, U., & Liede, S. (2004). Subtribal division of Ceropegieae (Apocynaceae-Asclepiadoideae). *Taxon, 53*(1), 61-72. doi:10.2307/4135489

Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology, 12*(11). doi:10.1186/gb-2011-12-11-r112

Moore, M. J., Bell, C. D., Soltis, P. S., & Soltis, D. E. (2007). Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences of the United States of America, 104*(49), 19363-19368. doi:10.1073/pnas.0708072104

Moore, M. J., Dhingra, A., Soltis, P. S., Shaw, R., Farmerie, W. G., Folta, K. M., & Soltis, D. E. (2006). Rapid and accurate pyrosequencing of angiosperm plastid genomes. *Bmc Plant Biology, 6*. doi:10.1186/1471-2229-6-17

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., . . . Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research, 39*(13). doi:10.1093/nar/gkr344

Nock, C. J., Waters, D. L. E., Edwards, M. A., Bowen, S. G., Rice, N., Cordeiro, G. M., & Henry, R. J. (2011). Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnology Journal, 9*(3), 328-333. doi:10.1111/j.1467-7652.2010.00558.x

Omlor, R. (1996). Notes on Marsdenieae (Asclepiadaceae): A new, unusual species of *Hoya* from Northern Borneo. *Novon, 6*, 288-294.

Parks, M., Liston, A., Cronn, R. (2010). Meeting the challenges of non-referenced genome assembly from short-read sequence data. *Acta Horticulturae, 859*: 323-332.

Palmer, J. D. (1991). Plastid Chromosomes: Structure and Evolution. In *The Molecular Biology of Plastids* (pp. 5-53). San Diego, CA: Academic Press, Inc.

Potgieter, K., & Albert, V. A. (2001). Phylogenetic relationships within Apocynaceae s.l. based on trn L intron and trn L-F spacer sequences and propagule characters. *Annals of the Missouri Botanical Garden, 88*, 523-549.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Rabinowicz, P. D., & Bennetzen, J. L. (2006). The maize genome as a model for efficient sequence analysis of large plant genomes. *Current Opinion in Plant Biology, 9*(2), 149-156. doi:10.1016/j.pbi.2006.01.015

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., . . . Yorke, J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research, 22*(3), 557-567. doi:10.1101/gr.131383.111

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research, 43*(6). doi:10.1093/nar/gku1341

Senbladd, B., & Bremer, B. (2002). Classification of Apocynaceae s.l. according to a new approach combining Linnaean and phylogenetic taxonomy. *Systematic Biology, 51*, 389-409.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31*(19), 3210-3212. doi:10.1093/bioinformatics/btv351

Simon, S. A., Zhai, J. X., Nandety, R. S., McCormick, K. P., Zeng, J., Mejia, D., & Meyers, B. C. (2009). Short-read sequencing technologies for transcriptional analyses. *Annual Review of Plant Biology, 60*, 305-333. doi:10.1146/annurev.arplant.043008.092032

Steele, P. R., Hertweck, K. L., Mayfield, D., McKain, M. R., Leebens-Mack, J., & Pires, J. C. (2012). Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. *American Journal of Botany, 99*(2), 330-348. doi:10.3732/ajb.1100491

Straub, S. (n.d.) [Raw reads for *Alyxia siamensis* and *Aspidosperma cruentum*]. Unpublished raw data.

Straub, S. (2012). *Botany 2012: Introduction to Next-Generation Sequencing Workshop Practical Exercises*. [PDF]. Retrieved from http://milkweedgenome.org/sites/default/files/workshopFiles/Botany_2012_NGS_workshop_exercises_0.pdf

Straub, S. C. K., Cronn, R. C., Edwards, C., Fishbein, M., & Liston, A. (2013). Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). *Genome Biology and Evolution, 5*(10), 1872-1885. doi:10.1093/gbe/evt140

Straub, S. C. K., Fishbein, M., Livshultz, T., Foster, Z., Parks, M., Weitemier, K., . . . Liston, A. (2011). Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *Bmc Genomics, 12*. doi:10.1186/1471-2164-12-211

Straub, S. C. K., Moore, M. J., Soltis, P. S., Soltis, D. E., Liston, A., & Livshultz, T. (2014). Phylogenetic signal detection from an ancient rapid radiation: Effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. *Molecular Phylogenetics and Evolution, 80*, 169-185. doi:10.1016/j.ympev.2014.07.020

Straub, S. C. K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., & Liston, A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany, 99*(2), 349-364. doi:10.3732/ajb.1100335

Verhoeven, R. L., Liede, S., & Endress, M. E. (2003). The tribal position of *Fockea* and *Cibirhiza* (Apocynaceae : Asclepiadoideae): evidence from pollinium structure and cpDNA sequence data. *Grana, 42*(2), 70-81. doi:10.1080/001731310310012549

Whittall, J. B., Syring, J., Parks, M., Buenrostro, J., Dick, C., Liston, A., & Cronn, R. (2010). Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology, 19*, 100-114. doi:10.1111/j.1365-294X.2009.04474.x

Wysocki, W. P., Clark, L. G., Kelchner, S. A., Burke, S. V., Pires, J. C., Edger, P. P., . . . Duvall, M. R. (2014). A multi-step comparison of short-read full plastome sequence assembly methods in grasses. *Taxon, 63*(4), 899-910.

Yue, F., Cui, L. Y., Depamphilis, C. W., Moret, B. M. E., & Tang, J. J. (2008). Gene rearrangement analysis and ancestral order inference from chloroplast genomes with inverted repeat. *Bmc Genomics, 9*. doi:10.1186/1471-2164-9-s1-s25

APPENDIX A

Results of Mann-Whitney U test comparing NG50 between *Asclepias syriaca* and *Nerium oleander* for Marsdenieae taxa alone and Marsdenieae plus *Alyxia siamensis* and *Aspidosperma cruentum*.

NG50 Marsdenieae taxa		**NG50 All Taxa**	
W	p-value	W	p-value
57.5	0.5955	69.5	0.9077

Results of Independent Samples t-tests comparing percent of contigs that aligned to each reference plastome and percent coverage of each reference plastome. Results are presented for Marsdenieae taxa alone and Marsdenieae plus *Alyxia siamensis* and *Aspidosperma cruentum*.

	t	df	p-value
Percent Contigs that Aligned, Marsdenieae Taxa	1.4589	16.53	0.1633
Percent Contigs that Aligned, All Taxa	0.23934	21.729	0.8131
Percent Coverage of Plastome by Aligned Contigs, Marsdenieae Taxa	1.5967	17.938	0.1278
Percent Coverage of Plastome by Aligned Contigs, All Taxa	0.23195	21.954	0.8187