USING COMPARATIVE PLASTOMICS TO IDENTIFY POTENTIALLY INFORMATIVE NON-CODING REGIONS FOR BASAL ANGIOSPERMS, WITH A FOCUS ON *ILLICIUM* (SCHISANDRACEAE)

By

Opal Rayne Leonard

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Science in Biology

Middle Tennessee State University
December 2015

Thesis Committee:

Dr. Ashley B. Morris, Chair

Dr. Rebecca Seipelt-Thiemann

Dr. Joey Shaw

Dedicated to Sam and Opal Blackwell.

ACKNOWLEDGEMENTS

I would like to extend gratitude to Joey Shaw, Rebecca Seipelt-Thiemann, Kurt Neubig, Mike Moore, and Sarah Bergemann for their endless supply of expertise and guidance. I would also like to thank Mark Weathington for his willingness to ship plant materials overnight, with only a moment's notice, and Dr. Zhiduan Chen and the members of his lab, particularly Miao Sun, Chen Min, and Jian Zhang, who made me feel welcome in Beijing. Special thanks goes to my adviser, Dr. Ashley B. Morris, and the members of her lab, whose guidance, feedback, and camaraderie made the process easier and enjoyable. And infinite gratitude to Christopher Davis and my family, whose support through this process was irreplaceable.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER 1: BACKGROUND AND OBJECTIVES

Chloroplasts are multifunctional plant organelles that possess their own genetic material and are generally maternally inherited, with exceptions such as *Pelargonium* (Birky, 1978) and some gymnosperms (Zhang et al, 2003). The chloroplast genome (hereafter referred to as the plastome) is between 115 and 165 kilobase pairs (kb) in length, with two inverted repeats and large and small single copy regions (Luo et al., 2014). Plastomes show a high degree of conservation in size, structure, gene content, and linear order of genes in land plants (Palmer, 1985; Shaw et al., 2007). Plastomes are also present in high numbers within cells and show a low rate of recombination, making them ideal for providing information about evolutionary relationships and divergence times among organisms. Overall, the plastome has become an important and useful tool for understanding the evolutionary history of angiosperms (Palmer et al., 1988; Ravi et al., 2008; Li et al., 2013). Here the utility of non-coding chloroplast DNA (NC-cpDNA) for low-level phylogenetic analyses is explored.

Non-coding chloroplast DNA (NC-cpDNA) has numerous applications in systematics and evolutionary biology, such as elucidating the origin of domesticated species, tracing biogeographic movements, and clarifying complex relationships among species (Taberlet et al., 1991; Kelchner, 2000; Sennblad and Bremer, 2000; Bremer et al., 2002). Non-coding chloroplast regions are proving to be useful for both high and low-level phylogenetics and phylogeography (Sarkinen and George, 2013). Non-coding regions of the plastome are being explored further for taxonomic studies under the

expectation that non-coding regions are under less selective constraint than coding regions and therefore have the potential to provide higher levels of variation for phylogenetic analyses at lower taxonomic levels (Shaw et al., 2005; Li et al., 2015).

Taberlet et al. (1991) encouraged increased use of non-coding chloroplast DNA sequences by developing universal primers for NC-cpDNA PCR amplification for use with both intra- and interspecific phylogenetic studies. Shaw et al. (2014) determined that while there are no universally informative NC-cpDNA regions, there are a few that are consistently informative across the angiosperms. Downie and Jansen (2015) found that certain NC-cpDNA regions, such as *trnH-psbA,* were highly variable in some groups in the Apiales, but not variable in others. Li et al. (2013) have done similar work with Araliaceae, using correlation analyses to reveal a positive linear relationship between percentages of parsimony informative sites and percentages of variable sites in candidate regions. Of the 25 variable non-coding regions found by Li et al. (2013) to be potentially useful phylogenetic markers, seven of them were among the top potentially informative regions found by Shaw et al. (2007) across all angiosperms. The overall use of potentially informative NC-cpDNA markers has not been thoroughly assessed in the basal angiosperms.

Basal angiosperms are a group of primitive, non-monocot, non-eudicot angiosperms considered to represent the earliest lineages of flowering plants (Soltis et al., 2009). The changes that lead to the progress and diversity of angiosperm lineages are of particular interest to basic and applied plant biologists and have garnered much attention for basal angiosperms in recent years (Doyle and Endress, 2000; Denk and Oh, 2005).

The diversity of basal angiosperms provides insight into angiosperm adaptation (Bliss et al., 2013). This research focuses on the basal angiosperms, with particular emphasis on *Illicium* (Schisandraceae).

The purpose of this investigation is to complete a comparative analysis of the whole plastomes of five members of *Illicium,* representing both Old and New World clades of *Illicium*, and to determine which NC-cpDNA regions are potentially most informative for phylogenetic analyses within this group. In addition, results of the comparison will be assessed within the broader context of basal angiosperms. This type of comparison has not been done with basal angiosperms as the focus, which makes this study a novel contribution to the literature. Newly sequenced *Illicium* plastomes resulting from this work will be added to the limited number of current publicly available plastome data. Furthermore, a simple, quick technique has been developed to aid researchers in the comparative plastomics analysis pipeline. Specific questions to be answered from this work include: How do the most variable NC-cpDNA regions differ across taxonomic levels in *Illicium*? How do the most variable NC-cpDNA regions for basal angiosperms differ from the most variable regions across all angiosperms? How do these results compare to those of similar comparative plastomics studies?

REFERENCES

Bliss, B. J., S. Wanke, A. Barakat, S. Ayyampalayam, N. Wickett, P. K. Wall, Y. Jiao, et al. 2013. Characterization of the basal angiosperm *Aristolochia fimbriata*: a potential experimental system for genetic studies. *BMC Plant Biology* 13:13.

Bock, R. 2014. Genetic engineering of the chloroplast: novel tools and new applications. *Current Opinion in Biotechnology* 26: 7-13.

Bremer, B., K. Bremer, N. Heidari, P. Erixon, R. G. Olmstead, A. A. Anderberg, M. Kallersjo, et al. 2002. Phylogenetics of asterids based on 3 coding and 3 non-coding chloroplast DNA markers and the utility of non-coding DNA at higher taxonomic levels. *Molecular Phylogenetics and Evolution* 24: 274-301.

Denk, T. and I. C. Oh. 2006. Phylogeny of Schisandraceae based on morphological data: evidence from modern plants and the fossil record. *Plant Systematics and Evolution* 256: 113-145.

Dong, W., J. Liu, J. Yu, l. Wang, and S. Zhou. 2012. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *Plos One* 7: e35071.

Downie, S. R., and R. K. Jansen. 2015. A comparative analysis of whole plastid genomes from the Apiales: expansion and contraction of the inverted repeat, mitochondrial to plastid transfer of DNA, and identification of highly divergent non-coding regions. *Systematic Botany* 40: 336-351.

Doyle, J. A., and P. K. Endress. 2000. Morphological phylogenetic analysis of basal angiosperms: comparison and combination with molecular data. *International Journal of Plant Sciences* 161: 121-153.

Kelchner, S. A. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. *Annals of the Missouri Botanical Garden* 87: 482-498.

Li, R., P.-F. Ma, J. Wen, and T.-S. Yi. 2013. Complete sequencing of five Araliaceae chloroplast genomes and the phylogenetic implications. *Plos One* 8: e78568.

Li, X., Y. Yang, R. J. Henry, M. Rossetto, Y. Wang, and S. Chen. 2015. Plant DNA barcoding: from gene to genome. *Biological Reviews* 90: 157-166.

Luo, J, B. W. Hou, Z. T. Niu, W. Liu, Q. Y. Xue, and X. Y. Ding. 2014. Comparative chloroplast genomes of photosynthetic orchids: insights into evolution of the Orchidaceae and development of molecular markers for phylogenetic applications. *Plos One* 9: e99016.

Palmer, J. D., R. K. Jansen, H. J. Michaels, M. W. Chase, and J. R. Manhart. 1988. Chloroplast DNA variation and plant phylogeny. *Annals of the Missouri Botanical Garden* 75: 1180-1206.

Ravi, V., J. P. Khurana, A. K. Tyagi, and P. Khurana. 2008. An update on chloroplast genomes. *Plant Systematics and Evolution* 271: 101-122.

Saerkinen, T., and M. George. 2013. Plastid marker variation: can complete plastid genomes from closely related species help? *Plos One* 8: e82266.

Sennblad, B., and B. Bremer. 2000. Is there a justification for differential *a priori* weighting in coding sequences? A case study from *rbcl* and Apocynaceae s.l. *Systematic Biology* 49: 101-113.

Shaw, J., E. B. Lickey, E. E. Schilling, and R. L. Small. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. *American Journal of Botany* 94: 275-288.

Shaw, J., H. L. Shafer, O. R. Leonard, M. J. Kovach, M. Schorr, and A. B. Morris. 2014. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. *American Journal of Botany* 101: 1987-2004.

Shaw, J., E. B. Lickey, J. T. Beck, S. B. Farmer, W. S. Liu, J. Miller, K. C. Siripun, et al. 2005. The tortoise and the hare II: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* 92: 142-166.

Soltis, P. S., S. F. Brockington, M.-J. Yoo, A. Piedrahita, M. Latvis, M. J. Moore, A. S. Chanderbali, et al. 2009. Floral variation and floral genetics in basal angiosperms. *American Journal of Botany* 96: 110-128.

Taberlet, P., L. Gielly, G. Pautou, and J. Bouvet. 1991. Universal primers for amplification of 3 noncoding regions of chloroplast DNA. *Plant Molecular Biology* 17: 1105-1109.

Zhang, Q., Y. Liu, and Sodmergen. 2003. Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. *Plant and Cell Physiology* 44: 941-951.

CHAPTER 2: USING COMPARATIVE PLASTOMICS TO IDENTIFY POTENTIALLY INFORMATIVE NON-CODING REGIONS

Abstract

- *Premise of study:* Comparative plastomics provides a method for choosing the most informative tools for a given study group, often a difficult process due to limited available data for targeted taxa. Many of the most commonly used chloroplast DNA regions for phylogenetic analyses are not the regions predicted to be most variable by pairwise taxonomic comparisons across varying study groups, and therefore may not be the most useful regions available, demonstrating the need for testing the top potential informative regions. This research seeks to add to the groundwork of basal angiosperm phylogenetics by providing an understanding of the tools available in this important group of flowering plants.
- *Methods:* A comparative analysis was completed using the whole plastomes of five members of *Illicium*: *I. oligandrum, I. henryi, I. cubense, I. floridanum,* and *I.ekmanii*. An additional analysis was completed using representatives of the broader basal angiosperms: *Amborella trichopoda, Nymphaea alba, Nymphaea mexicana, Nuphar advena,* and *Trithuria inconspicua.* Perl scripts were written to expedite the comparative screening analysis. In each case, the objective was to identify the most potentially variable non-coding chloroplast DNA regions for phylogeny reconstruction and phylogeographic analyses.

- *Key results:* The most variable regions identified for *Illicium* were *petN-psbM, rpl32-trnL, cemA-petA, petB intron,* and *psaC-ndhE.* The most variable regions across the basal angiosperms were *psbE-petL, rpoB-trnC, matK, trnE-trnT,* and *psbM-trnD.* Four regions, *ndhF-rpl32, ndhC-trnV, rps16-trnQ,* and *trnT-psbD*, were listed as top performers in a previous study, but were unable to be sequenced in *Illicium* and were excluded.
- *Conclusions:* The most variable regions differed between different taxonomic levels in the *Illicium* and basal angiosperm comparative analyses. The *Illicium* regions that did not amplify and were therefore excluded may be the most variable regions in *Illicium,* and warrant further testing. While a few regions stand out as variable in all analyses at lower taxonomic levels, there are clear differences in which regions will likely be phylogenetically informative at different taxonomic scales. Therefore, researchers who choose not to use next-generation sequencing methods should employ a screening process in the group of interest before beginning a phylogenetic analysis.

Introduction

Since the first chloroplast gene was suggested for use for phylogenetic analysis (*rbcL*), chloroplast DNA has become increasingly utilized for phylogeny reconstruction (Chase et al., 1993; Olmstead and Palmer, 1994; Sennblad and Bremer, 2000; Shaw et al., 2005; Hansen et al., 2007; Dong et al., 2012; Ruhfel et al., 2014). The conservative nature of the chloroplast genome (i.e., plastome) has made it a valuable molecule for phylogenetic studies at all levels, such that chloroplast DNA sequences are currently the most common source of data for construing plant phylogenies (Shaw et al., 2005; Li et al., 2015). However, the same conservative nature that makes chloroplast coding regions useful for deeper phylogenetic reconstruction has been assumed to make them less useful for inter- and intraspecific phylogenetic studies (Palmer et al., 1988; Taberlet et al., 1991). At the shallowest taxonomic level, it can be difficult to find enough genetic variability within chloroplast coding regions to establish a robust phylogenetic hypothesis. Researchers have suggested using non-coding regions of the plastome, in addition to the traditionally used coding regions to correct for the conservative nature of chloroplast genes (Doyle, 2013). Phylogenetic relationships that have remained unresolved due to low rates of nucleotide substitutions in markers may be resolved using non-coding regions that are more rapidly evolving (Li et al., 2013).

Shaw et al. (2014) found that the most variable non-coding DNA regions of the chloroplast are not those currently used in most phylogeographic studies, and determined that while there are no universally most informative NC-cpDNA regions, there are several that are consistently informative across the angiosperms. Three large plastome

regions were found to be consistently variable: the area from *ccsA* to *ndhF* (containing *ndhF-rpl32* and *rpl32-trnL),* the area from *matK* to *3'trnG* (containing *matK, 5'trnK-3'rps16, 5'rps16-trnQ,* and *trnS-5'trnG),* and the area between *rpoB* to *psbD.* Within these regions, *ndhF-rpl32, rpl32-trnL, ndhC-trnV,* and *5'rps16-trnQ* were most informative overall (refer to Figure 1 for a representation of gene content and order in the *Illicium oligandrum* plastome). That is not to say that these areas are most variable for all angiosperms, but it was recommended that screening for the most informative regions in a given group would begin with the most informative regions overall in absence of a comparative plastomic approach (Shaw et al., 2014).

As next-generation sequencing (NGS) has become more accessible to researchers, previously unexplored regions of the plastome have become obvious candidates for phylogenetics (Shaw et al., 2007). The accessibility of NGS for a given lab often depends on funding, computational support, and technical expertise within the lab. These limitations can render NGS inaccessible to researchers at smaller or primarily undergraduate institutions. Compared to past costs of whole genome sequencing, NGS is relatively cheap and accessible for many researchers (Godden et al., 2012). However, NGS is cheapest when library building and sequencing is outsourced and performed in bulk; this means that the per-plastome cost is relatively low, but the overall cost of the sequencing job can be high. Furthermore, a large amount of DNA is required for NGS, which is a limiting factor for researchers working with problematic plant taxa due to secondary chemistry or poor preservation.

As plastome sequencing becomes more common, it is important that researchers have access to the best tools available for raw data analysis and assembly. However, even

as genome sequencing has gotten easier, genome annotation has arguably become more challenging (Yandell and Ence, 2012). There is a steep learning curve associated with the raw data obtained from NGS, which requires computational power and trained expertise in order to be efficiently and correctly assembled. There are many options available that facilitate assembly of NGS data, but most programs are not user friendly and require at least a basic knowledge of computer science. There are few step-by-step tutorials available, so researchers must rely on training from other experienced researchers – something that may be difficult at an institution previously lacking the equipment and funding to perform NGS. In addition, each genome sequencing and assembly project is likely to have unique issues that must be resolved, so no singular pipeline can be used for all analyses.

As a result, many researchers make a choice between NGS and Sanger sequencing. For some, it is most cost-effective to sequence a few plastomes using NGS. Two to three plastomes are useful for screening NC-cpDNA regions for variability so that Sanger sequencing can be used for subsequent data collection of the most informative regions for a given group (Li et al., 2015). Therefore, further exploring the utility of non-coding DNA sequences and elucidating the most potentially informative non-coding regions will be useful for those choosing Sanger sequencing for their research (Shaw et al., 2014). The current research focuses on variability in *Illicium* NC-cpDNA regions, set within the context of the basal angiosperms.

Basal angiosperms are important to our broader understanding of flowering plants because they provide insight into the diversity within angiosperms, polarize analyses of flowering plant evolution, and make functional inferences about the common ancestor of

early angiosperms (Parkinson et al., 1999; Bliss et al., 2013). Basal angiosperms consist of the orders of the ANA grade, containing Amborellales, Nympheales, and Austrobaileyales (Doyle, 2000; Hilu et al., 2003; Chien et al., 2011). The basal grade of angiosperms lacks phylogenetic support; for example, Amborellaceae, Nymphaeales, and Austrobaileyales are consistently placed as sister taxa to all other angiosperms, but branching order has been disputed, especially between Amborellaceae and Nymphaeales (Jansen et al., 2007; Drew et al., 2014). Even though molecular data have provided the greatest resolution for this group to date, data sets tend to have too few characters, and additional clarification is needed (Soltis et al., 2009).

This work has a particular emphasis on *Illicium,* a genus in the family Schisandraceae, within the order Austrobaileyales. APG III does not recognize *Illicium* as a family unto itself, but instead only recognizes Schisandraceae, which includes *Illicium* (Stevens, 2015). However, several sources treat *Illicium* as the sole genus in the family Illiciaceae, and Illiciaceae as sister group to Schisandraceae (Smith, 1947; Hao et al., 2000; Morris et al., 2007; Soltis et al., 2009). *Illicium,* commonly known as star anise, is a monophyletic group of basal angiosperms comprised of 30-40 species in southeastern North America, Mexico, Greater Antilles, and East to Southeast Asia. *Illicium* is an economically and medicinally important plant in Southeast Asia and around the world. It is most easily recognized by its star-shaped fruits, which give the group its common name. Star anise is used in many dishes in Chinese and Indian cuisine, and *Illicium* is very commonly used in Chinese folk medicine (Meizi et al., 2012). *Illicium verum* is also of medicinal importance to Western medicine due to an abundance of shikimic acid, an important precursor to the active ingredient in the anti-viral medication Tamiflu (Ward et

al., 2005; Awang, 2006; Avula, 2009; Techen et al., 2008). *Illicium* is unique among basal angiosperms, with some species having larger ethereal oil cells than any other ANA grade family, which have been used as taxonomic characters in previous studies (Carpenter, 2006).

As *Illicium* is a group of early diverging angiosperms that exhibits the well-documented floral disjunction between the New World and Old World (Figure 2), it is well positioned for studies of biogeography, floral development, and molecular evolution (Morris et al., 2007). However, taxonomic identification has historically been difficult because many of the recognized species are morphologically similar (Smith, 1947). Smith (1947) recognized that *Illicium* species delimitation is difficult due to homoplasy among morphological characters commonly used for taxonomic differentiation. *Illicium* is an ideal model group for testing the viability of screening whole plastomes and the utility of NC-cpDNA due to the small size of the group, the noteworthy features such as biogeographical disjunction, and historical difficulties with taxonomic differentiation. Prior to this study, there was little data available for studies of *Illicium* or sister taxa; there was one *Illicium* whole plastome available in GenBank, and there were no plastomes representing the *Illicium* sister taxa, *Schisandra* and *Kadsura.* This study adds to the current data available for researchers, and establishes variable markers for future projects through a plastome screening analysis.

Methods and Materials

Data set – A total of 10 plastomes were used for this study. The data set included four *Illicium* plastomes sequenced for this research: *I. cubense* (SRX1317965)*, I. ekmanii* (SRX1317968)*, I. floridanum* (SRX1317966)*, I. henryi* (SRX1317964) (Austrobaileyales) (Table 1). Six additional basal angiosperm plastomes available in NCBI Organelle Genome Resources were also included: *Amborella trichopoda* (NC_005086; Amborellales); *Trithuria inconspicua* (NC_020372)*, Nymphaea alba* (NC_006050)*, Nymphaea mexicana* (NC_024542)*,* and *Nuphar advena* (NC_008788) all from Nympheales; and *Illicium oligandrum* (NC_009600; Austrobaileyales) (Table 1). In total, these plastomes represent the three orders of the ANA grade of basal angiosperms (Doyle, 2000; Hilu et al., 2003; Chien et al., 2011).

Sequencing **–** Total genomic DNA was extracted from each specimen following a modified CTAB protocol according to Neubig et al. (2014). Samples were shipped to RAPiD Genomics for sequencing (Gainesville, Florida). DNA was quantified using dye intercalating PicoGreen reagent. Samples were sheared to average fragment size of 350 bp. Illumina TruSeq-like libraries were built with 29 unique eight bp indexes, and pooled. Samples were sequenced on an Illumina HiSeq 2500 with 100 bp single-end reads of nuclear, mitochondrial, and chloroplast DNA. Raw sequence data is available in NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra).

Assembly of new plastomes – Sequencing artifacts were trimmed from the raw data via Trimmomatic (Usadel Lab, Aachen University, Aachen, Germany) using the TruSeq3 adapter file. Raw reads for were checked for quality using FastQC (Babraham Informatics, Cambridge, UK). Reads were assembled with the map to reference function in Geneious (v.7, Biomatters Inc., San Francisco, CA, USA) using *Illicium oligandrum* (NC_009600) as the reference sequence on medium-low sensitivity. Initial assemblies were completed after removing inverted repeat region B from the reference sequence. After assembly, the remaining inverted repeat region in each plastome was extracted, converted to the reverse complement, and copied back into the appropriate position. Each assembly was annotated using the Geneious annotation function to copy annotations from the reference sequence. The assemblies were manually checked for sequencing artifacts and ambiguities.

Comparative plastomics analysis –Four comparisons of NC-cpDNA regions representing different taxonomic scales were completed for this study (Table 2): 1) within the Old World clade of *Illicium* (two species); 2) within the New World clade of *Illicium* (three species); 3) across all of *Illicium* (five species); 4) across the basal angiosperms (10 species representing four families). Regions were selected based on the regions surveyed in Shaw et al. (2014) to create a comparable data set. Regions shorter than 100 bp were excluded, due to the likelihood of recovering few potentially informative characters (PICs) for the effort expended.

For each comparison, each individual NC-cpDNA region was aligned across accessions and scored manually for potentially informative characters (PICs) in order to determine the potentially most informative non-coding regions within *Illicium* species (Shaw et al., 2007). Each region to be analyzed was extracted and aligned across accessions using the MUSCLE alignment software supported within Geneious (version 7) at default parameters. A list of regions utilized in each analysis is included in Appendix A. The PICs included substitutions, insertions or deletions (indels), and inversions, with each substitution or inversion scored as a single character. Indels were scored as gaps following Simmons and Ochoterena (2000). Indels resulting from length variation in mononucleotide repeats were not included. The total number of PICs in each region was used to find the normalized PIC value, which represents the percentage contribution of each NC-cpDNA region to the overall variability in a lineage (Shaw et al., 2014). This is found by dividing the number of PICs in a region by the overall number of PICs in the entire comparison.

$$normalized\ PIC\ value = \frac{number\ of\ PICs\ in\ region}{number\ of\ PICs\ in\ the\ total\ comparison}$$

This normalized value takes into account different evolutionary rates between comparisons and reduces the possibility of overrepresentation of a comparison with a higher number of PICs (Shaw et al, 2007). This allows for comparison of NC-cpDNA regions between lineages. It must be noted that when using the normalized PIC value, plastome comparisons with very few PICs overall must be eliminated due to the NC-

cpDNA regions containing those PICs having very high normalized PIC values (Shaw et al., 2014).

This comparison is often completed manually by obtaining a multiple alignment for each NC-cpDNA region and counting the PICs by eye. However, this study involved many comparisons across multiple taxonomic levels and would have been labor intensive if completed manually. Therefore, a series of simple Perl scripts, called PIC Counter, were developed for this project, which allowed multiple fasta-formatted alignments to be automatically scored for substitutions (Leonard et al., in prep). These scripts scored both substitutions and indels for pairwise comparisons, but three-way and 10-way comparisons were manually scored for indels following Simmons and Ochoterena (2000).

The ten-way analysis of basal angiosperms was conducted differently due to the variable nature of NC-cpDNA regions and the inability to align some of these regions at higher taxonomic levels. The common barcodes *matK* and *rbcL* were extracted, aligned, and scored across all ten basal angiosperm plastomes to provide a baseline of variability at the ordinal taxonomic level. Since some of the 10-way multiple alignments were too variable to reliably align, the alignments were manually checked, and regions with alignments that were unreliable due to high variability were excluded. Alignments for which *Illicium* data were missing were excluded from all comparisons.

Basal angiosperm literature review – A literature was completed to document which chloroplast DNA regions have been used in phylogeny reconstruction of basal

angiosperms. A search was performed in Web of Science, limiting the time period to 2010-2015, using search terms "phylogeny or phylogeography," "chloroplast or plastid or cpDNA," and each family included in the basal angiosperm orders, Amborellales (Amborellaceae), Nympheales (Nympheaceae, Hydatellaceae, Cabombaceae), and Austrobaileyales (Schisandraceae, Austrobaileyaceae, Trimeniaceae). Web of Science allows several search fields to be used at once, so all search terms were entered concurrently. In that way, seven searches were completed, each with all search terms and a basal angiosperm family name. These parameters yielded 16 publications. These papers were reviewed for information including: author, year published, family, markers used, length, PIC (if reported), and sequencing methods.

Results

Plastome size, content, and organization – Illicium plastome sizes ranged from 147,467 to 148,187 base pairs (bp), all smaller than *I. oligandrum*, at 148,552 bp (Table 3). The amount of missing data relative to the reference plastome ranges from 205 bp to 547 bp. The plastomes were AT rich, with AT content at 60.85% for each. The number of reads mapped to the reference ranged from 36,151 to 118,595. The average read sizes were 99 bp for each plastome. Mean depth of coverage ranged from 27 to 88 reads, with maximum depth from 100 to 244 reads. The number of reads mapped affects the depth of coverage of an assembled plastome, and does not affect the size of the consensus

sequence. The total number of genes annotated in each plastome was 126. Each plastome shared the 10 kb IR contraction originally reported in *I. oligandrum* (Hansen et al., 2007).

Regions not sequenced – Several regions in the *Illicium* plastomes failed to map to the *I. oligandrum* reference: *ndhC-trnV, ndhF-rpl32, rps16-trnQ, trnT-psbD, atpF-atpH, petN-psbM, rps8-rpl14, trnL-ndhF, trnS-trnG, trnT-trnL, ycf2-trnL* (Table 4). The potential reasons these regions failed to map to the reference sequence include: the regions did not map to the reference sequence due to high species divergence; or the regions did not amplify during the sequencing process. These regions are potentially AT rich. Troubleshooting was employed to verify that these regions were not present in the raw reads. Troubleshooting involved using the contigs surrounding these regions in an attempt to grow the ends of the contigs using reference-guided assembly, namely, using the contigs as the reference. The process can be repeated iteratively to continue growing the contig. However, troubleshooting did not yield any significant decreases in missing data. Primers have been developed via Primer3 (version 2.3.4 -- http://bioinfo.ut.ee/primer3/) to target these missing regions for Sanger sequencing in a future study (Table 5).

Screening for potentially informative characters – The number of regions surveyed differs between analyses. In the *Illicium* comparisons, this is a result of exclusion of regions that were not sequenced in the plastomes. In the basal angiosperm comparison, the same regions excluded for *Illicium* apply, in addition to the exclusion of regions too variable to align at higher taxonomic levels. The most variable NC-cpDNA regions

identified in this study were ranked by normalized PIC value (Table 6). It must be noted that the *Illicium* regions that did not amplify and were therefore excluded may be the most variable regions in *Illicium,* and warrant further testing. The most variable of the 81 regions surveyed for the New World *Illicium* comparison were, in order of highest normalized PIC value: *petN-psbM, rpl32-trnL, cemA-petA, psbM-trnD, trnM-atpE, trnQ-psbK, matK-rps16, psbK-psbI, atpH-atpI,* and *trnH-psbA.* The most variable of the 89 regions surveyed for the Old World *Illicium* comparison were, in order of highest normalized PIC value: *petN-psbM, rps16-trnQ, petA-psbJ, trnS-trnG, cemA-petA, petB* intron*, trnT-psbD, ndhC-trnV, matK-rps16,* and *atpH-atpI.* The most variable of the 84 regions surveyed for the comparison across all available *Illicium* plastomes were, in order of highest normalized PIC value: *petN-psbM, rpl32-trnL, cemA-petA, petB* intron*, psaC-ndhE, trnQ-psbK, psbM-trnD, trnT-psbD, trnM-atpE,* and *matK-rps16.* The most variable of the 73 regions surveyed for the comparison across all available basal angiosperm plastomes were, in order of highest normalized PIC value: *psbE-petL, rpoB-trnC, matK, trnE-trnT, psbM-trnD, trnC-petN, rpl16* intron*, ycf3-trnS, trnF-ndhJ,* and *accd-psaI* (Table 7). Regions *petN-psbM, matK-rps16, atpH-atpI, ndhA* intron*, petD-rpoA,* and *rpl32-trnL* were excluded from the basal angiosperm (10 taxon) comparison due to high variability resulting in unreliable alignments (Table 7). Raw data, including regions surveyed, total number of PICs, length of each alignment, normalized PIC values, and percent variability are provided in Appendix A.

Basal angiosperm literature review – Recent publications involving phylogenetic studies of basal angiosperm families were surveyed. Of the 16 papers surveyed, 13 papers utilized chloroplast DNA markers, and of those, six publications used non-coding chloroplast DNA. Nine publications focused solely on a family within the basal angiosperms, while seven publications involved large-scale analyses across many plant taxa. Only four papers reported PICs for the analyses. Three papers utilized next-generation sequencing and 10 papers utilized Sanger sequencing.

Discussion

The objective of this study was to complete a comparative analysis of variation in non-coding chloroplast DNA regions at different taxonomic levels in *Illicium* and the broader basal angiosperms. The results generated here are based on those portions of the genome that we were able to obtain from NGS, including 99.7% of the genome (see Table 4 for a list of the 11 regions that were not obtained). The results indicate that the most variable regions differ at different taxonomic levels, as well as within clades within the same genus. While this trend is to be expected between orders, families or even genera, the results of the present study show that a sliding scale of plastid utility applies even among clades within genera. This work underscores the value of comparative plastomics as a tool for marker selection in both phylogenetics and phylogeoraphy.

Difficulties assembling and amplifying variable regions – In each *Illicium* plastome, some of the most variable regions as predicted by Shaw et al. (2014) would not map to the reference during assembly (Tables 4 and 5). It was inferred after extensive troubleshooting that the variable regions were not mapping to the reference sequence because the regions were not sequenced successfully. *Illicium* chloroplast DNA varies little between species, therefore the possibility that the regions were not mapping due to high divergence from the reference sequence was ruled out. However, it is a possibility that the regions not sequenced are the most variable and potentially most informative for *Illicium*, inferred due to the high variability found in analyses in Shaw et al. (2014). Therefore, future studies will involve the incorporation of those regions into a comparative analysis. The amount of data missing from each plastome is listed as a percentage in Table 4.

A sliding scale of variability – In 67% of regions surveyed in the *Illicium* comparisons, the normalized PIC value was less than one percent, indicating that coding regions and many non-coding regions are not variable enough for species delimitation (Appendix 1). However, there were several non-coding regions not previously used in *Illicium* with sufficient variation to be potentially informative. Some of these same variable regions were too variable to be aligned across all of the basal angiosperm plastomes included in the analysis (Tables 6 and 7). The regions deemed to be potentially informative at the highest taxonomic level screened in this research were not particularly variable at the inter-specific level in *Illicium*. It should be noted that certain non-coding regions do

appear to be potentially informative at higher taxonomic levels, in spite of high variability. There are alignable non-coding regions that appear to be more variable than the common barcodes: *petN-psbM, matK-rps16, psbE-petL, atpH-atpI,* and *rpob-trnC.* This indicates that a sliding scale of variability may be established, and further indicates that a screening process should be used at the taxonomic level in question for a study.

Two studies involving *Illicium* identification via barcoding by Meizi et al. (2012) and Zhang et al. (2015) demonstrated that *psbA-trnH* had sufficient discriminating power among *Illicium* species. In the present study, *psbA-trnH* ranks 12th in the overall *Illicium* comparison, indicating that other, more variable markers, such as *petN-psbM, matK-rps16, cemA-petA,* and *trnT-psbD* (Figure 3), may be better barcodes for differentiating *Illicium* species that are morphologically similar. As shown by the present research, there is still no universally best region even within a target genus, due to differences in variability across taxonomic levels, and plastome screening in different groups and at different taxonomic levels yields overlapping, yet different results.

Further implications – Comparative plastomics has drawn attention to NC-cpDNA regions that were previously ignored as potentially phylogenetically informative markers (Shaw et al, 2014). Recent publications have indicated that the most variable, and potentially informative, non-coding regions differ between lineages (Sarkinen and George, 2013). Non-coding markers that are consistently variable across many lineages have also been found, such as *rpl32-trnL, ndhF-rpl32, rps16-trnQ,* and *trnT-psbD* (Sarkinen and George, 2013; Shaw et al, 2014; Downie and Jansen, 2015). However, of

all the groups surveyed in comparative analyses, the present study is the first to focus on the basal angiosperms, and the first to include a comparison at the family level. The results indicate that though non-coding regions are not utilized in studies involving many plant families, they may be useful for such a purpose. Non-coding regions have the potential to clarify relationships that are disputed, and should be included in the screening process. Comparative plastomics will continue to be a valuable tool for researchers as they determine which markers will yield the greatest resolution for the taxonomic groups and questions at hand.

Future directions – Future work will involve sequencing the regions in the *Illicium* plastomes that were not sequenced during next-generation sequencing for the current study. Those regions will then be included in the comparative analysis in *Illicium* and results reassessed. Furthermore, DNA sequences will be obtained for the remaining members in New World *Illicium,* either by Sanger sequencing or next-generation sequencing, in order to test the utility of the top most potentially informative regions.

REFERENCES

Avula, B., Y. H. Wang, T. J. Smillie, and I. A. Khan. 2009. Determination of shikimic acid in fruits of *Illicium* species and various other plant samples by LC-UV and LC-ESI-MS. *Chromatographic* 69: 307-314.

Awang, D. V. C. and D. Blumenthal. 2006. Tamiflu and star anise: securing adequate supplies of the oral antiviral for avian flu treatment. *Herbalgram* 70: 58-60.

Biswal, D. K., M. Debnath, S. Kumar, and P. Tandon. 2012. Phylogenetic reconstruction in the order Nymphaeales: ITS2 secondary structure analysis and in silico testing of maturase k (*matk*) as a potential marker for DNA bar coding. *BMC Bioinformatics* 13:17.

Bliss, B. J., S. Wanke, A. Barakat, S. Ayyampalayam, N. Wickett, P. K. Wall, Y. Jiao, et al. 2013. Characterization of the basal angiosperm *Aristolochia fimbriata*: a potential experimental system for genetic studies. *BMC Plant Biology* 13:13.

Bock, R. 2014. Genetic engineering of the chloroplast: novel tools and new applications. *Current Opinion in Biotechnology* 26: 7-13.

Borsch, T., J. H. Wiersema, C. B. Hellquist, C. Loehne, and K. Govers. 2014. Speciation in North American water lilies: evidence for the hybrid origin of the newly discovered Canadian endemic *Nymphaea loriana* sp nov (Nymphaeaceae) in a past contact zone. *Botany* 92: 867-882.

Bremer, B., K. Bremer, N. Heidari, P. Erixon, R. G. Olmstead, A. A. Anderberg, M. Kallersjo, et al. 2002. Phylogenetics of asterids based on 3 coding and 3 non-coding chloroplast DNA markers and the utility of non-coding DNA at higher taxonomic levels. *Molecular Phylogenetics and Evolution* 24: 274-301.

Carpenter, K. J. 2006. Specialized structures in the leaf epidermis of basal angiosperms: morphology, distribution, and homology. *American Journal of Botany* 93: 665-681.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, et al. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcl. Annals of Missouri Botanical Garden* 80: 528-580.

Chien, C. T.., S. Y. Chen, J. M. Baskin, and C. C. Baskin. 2011. Morphophysiological dormancy in seeds of the ANA grade angiosperm *Schisandra arisanensis* (Schisandraceae). *Plant Species Biology* 26: 99-104.

Denk, T. and I. C. Oh. 2006. Phylogeny of Schisandraceae based on morphological data: evidence from modern plants and the fossil record. *Plant Systematics and Evolution* 256: 113-145.

Dong, W., J. Liu, J. Yu, l. Wang, and S. Zhou. 2012. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *Plos One* 7: e35071.

Downie, S. R., and R. K. Jansen. 2015. A comparative analysis of whole plastid genomes from the Apiales: expansion and contraction of the inverted repeat, mitochondrial to plastid transfer of DNA, and identification of highly divergent non-coding regions. *Systematic Botany* 40: 336-351.

Doyle, J. A., and P. K. Endress. 2000. Morphological phylogenetic analysis of basal angiosperms: comparison and combination with molecular data. *International Journal of Plant Sciences* 161: 121-153.

Doyle, J. J. 2013. The promise of genomics for a "next generation" of advances in higher-level legume molecular systematics. *South African Journal of Botany* 89: 10-18.

Drew, B. T., B. R. Ruhfel, S. A. Smith, M. J. Moore, B. G. Briggs, M. A. Gitzendanner, P. S. Soltis, et al. 2014. Another look at the root of the angiosperms reveals a familiar tale. *Systematic Biology* 63: 368-382.

Fan, J.H., L. B. Thien, and Y. B. Luo. 2011. Pollination systems, biogeography, and divergence times of three allopatric species of *Schisandra* in North America, China, and Japan. *Journal of Systematics and Evolution* 49: 330-338.

Godden, G. T., I. E. Jordon-Thaden, S. Chamala, A. A. Crowl, N. Garcia, C. C. Germain-aubrey, J. M. Heaney, et al. 2012. Making next-generation sequencing work for you: approaches and practical considerations for marker development and phylogenetics. *Plant Ecology & Diversity* 5: 427-450.

Goremykin, V. V., S. V. Nikiforova, P. J. Biggs, B. Zhong, P. Delange, W. Martin, S. Woetzel, et al. 2013. The evolutionary root of flowering plants. *Systematic Biology* 62: 50-61.

Hansen, D. R., S. G. Dastidar, Z. Cai, C. Penaflor, J. V. Kuehl, J. L. Boore, and R. K. Jansen. 2007. Phylogenetic and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: *Buxus* (Buxaceae), *Chloranthus* (Chloranthaceae), *Dioscorea* (Dioscoreaceae), and *Illicium* (Schisandraceae). *Molecular Phylogenetics and Evolution* 45: 547-563.

Hao, G., R. M. K. Saunders, and M. L. Chye. 2000. A phylogenetic analysis of the Illiciaceae based on sequences of internal transcribed spacers (ITS) of nuclear ribosomal DNA. *Plant Systematics and Evolution* 223: 81-90.

Hilu, K. W., T. Borsch, K. Muller, D. E. Soltis, P. S. Soltis, V. Savolainen, M. W. Chase, et al. 2003. Angiosperm phylogeny based on *matk* sequence information. *American Journal of Botany* 90: 1758-1776.

Iles, W. J. D., P. J. Rudall, D. D. Sokoloff, M. V. Remizowa, T. D. Macfarlane, M. D. Logacheva, and S. W. Graham. 2012. Molecular phylogenetics of Hydatellaceae (Nymphaeales): sexual-system homoplasy and a new sectional classification. *American Journal of Botany* 99: 663-676.

Iles, W. J. D., C. Lee, D. D. Sokoloff, M. V. Remizowa, S. R. Yadav, M. D. Barrett, R. L. Barrett, et al. 2014. Reconstructing the age and historical biogeography of the ancient flowering-plant family Hydatellaceae (Nymphaeales). *BMC Evolutionary Biology* 14:102.

Jansen, R. K., Z. Cai, L. A. Raubeson, H. Daniell, C. W. Depamphilis, J. Leebens-Mack, K. F. Mueller, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences of the United States of America* 104: 19369-19374.

Kelchner, S. A. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. *Annals of the Missouri Botanical Garden* 87: 482-498.

Kim, C., J. Jung, H. R. Na, S. W. Kim, W. Li, Y. Kadono, H. Shin, et al. 2012a. Population genetic structure of the endangered *Brasenia schreberi* in South Korea based on nuclear ribosomal spacer and chloroplast DNA sequences. *Journal of Plant Biology* 55: 81-91.

Kim, J. S., H.W. Jang, J.S. Kim, H.J. Kim, and J.H. Kim. 2012b. Molecular identification of Schisandra chinensis and its allied species using multiplex PCR based on SNPs. *Genes & Genomics* 34: 283-290.

Li, R., P.-F. Ma, J. Wen, and T.-S. Yi. 2013. Complete sequencing of five Araliaceae chloroplast genomes and the phylogenetic implications. *Plos One* 8: e78568.

Li, X., Y. Yang, R. J. Henry, M. Rossetto, Y. Wang, and S. Chen. 2015. Plant DNA barcoding: from gene to genome. *Biological Reviews* 90: 157-166.

Luo, J, B. W. Hou, Z. T. Niu, W. Liu, Q. Y. Xue, and X. Y. Ding. 2014. Comparative chloroplast genomes of photosynthetic orchids: insights into evolution of the Orchidaceae and development of molecular markers for phylogenetic applications. *Plos One* 9: e99016.

Maia, V. H., M. A. Gitzendanner, P. S. Soltis, G. K.-S. Wong, and D. E. Soltis. 2014. Angiosperm phylogeny based on 18s/26s rDNA sequence data: constructing a large data set using next-generation sequence data. *International Journal of Plant Sciences* 175: 613-650.

Meizi, L., Y. Hui, L. Kun, M. Pei, Z. Wenbin, and L. Ping. 2012. Authentication of *Illicium verum* using a DNA barcode *psbA-trnH. Journal of Medicinal Plants Research* 6: 3156-3161.

Moore, M. J., P. S. Soltis, C. D. Bell, J. G. Burleigh, and D. E. Soltis. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences of the United States of America* 107: 4623-4628.

Moore, M. J., N. Hassan, M. A. Gitzendanner, R. A. Bruenn, M. Croley, A. Vandeventer, J. W. Horn, et al. 2011. Phylogenetic analysis of the plastid inverted repeat for 244 species: insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. *International Journal of Plant Sciences* 172: 541-558.

Morris, A. B., C. D. Bell, J. W. Clayton, W. S. Judd, D. E. Soltis, and P. S. Soltis. 2007. Phylogeny and divergence time estimation in *Illicium* with implications for New World biogeography. *Systematic botany* 32: 236-249.

Morton, C. M. 2011. Newly sequenced nuclear gene (XDH) for inferring angiosperm phylogeny. *Annals of the Missouri Botanical Garden* 98: 63-89.

Olmstead, R. G., and J. D. Palmer. 1994. Chloroplast DNA systematics - a review of methods and data-analysis. *American Journal of Botany* 81: 1205-1224.

Palmer, J. D. 1985. Comparative organization of chloroplast genomes. *Annual Review of Genetics* 19: 325-354.

Palmer, J. D., R. K. Jansen, H. J. Michaels, M. W. Chase, and J. R. Manhart. 1988. Chloroplast DNA variation and plant phylogeny. *Annals of the Missouri Botanical Garden* 75: 1180-1206.

Parkinson, C. L., K. L. Adams, and J. D. Palmer. 1999. Multigene analyses identify the three earliest lineages of extant flowering plants. *Current Biology* 9: 1485-1488.

Qiu, Y. L., L. B. Li, B. Wang, J. Y. Xue, T. A. Hendry, R. Q. Li, J. W. Brown, et al. 2010. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. *Journal of Systematics and Evolution* 48: 391-425.

Ravi, v., J. P. Khurana, A. K. Tyagi, and P. Khurana. 2008. An update on chloroplast genomes. *Plant Systematics and Evolution* 271: 101-122.

Ruhfel, B. R., M. A. Gitzendanner, P. S. Soltis, D. E. Soltis, and J. G. Burleigh. 2014. From algae to angiosperms: inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology* 14:23.

Saerkinen, T., and M. George. 2013. Plastid marker variation: can complete plastid genomes from closely related species help? *Plos One* 8: e82266.

Sennblad, B., and B. Bremer. 2000. Is there a justification for differential *a priori* weighting in coding sequences? A case study from *rbcl* and Apocynaceae s.l. *Systematic Biology* 49: 101-113.

Shaw, J., E. B. Lickey, E. E. Schilling, and R. L. Small. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. *American Journal of Botany* 94: 275-288.

Shaw, J., H. L. Shafer, O. R. Leonard, M. J. Kovach, M. Schorr, and A. B. Morris. 2014. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. *American Journal of Botany* 101: 1987-2004.

Shaw, J., E. B. Lickey, J. T. Beck, S. B. Farmer, W. S. Liu, J. Miller, K. C. Siripun, et al. 2005. The tortoise and the hare II: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* 92: 142-166.

Simmons, M. P. And H. Ochoterena. 2000. Gaps as characters in sequence-based phylogenetic analyses. *Systematic biology* 49: 369-381.

Smith, A. C. 1947. The families Illiciaceae and Schisandraceae. *Sargentia: a Continuation of the Contributions from The Arnold Arboretum Of Harvard University* 7:1-224.

Soltis, D. E., M. A. Gitzendanner, G. Stull, M. Chester, A. Chanderbali, S. Chamala, I. Jordon-Thaden, et al. 2013. The potential of genomics in plant systematics. *Taxon* 62: 886-898.

Soltis, D. E., S. A. Smith, N. Cellinese, K. J. Wurdack, D. C. Tank, S. F. Brockington, N. F. Refulio-Rodriguez, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98: 704-730.

Soltis, P. S., S. F. Brockington, M.-J. Yoo, A. Piedrahita, M. Latvis, M. J. Moore, A. S. Chanderbali, et al. 2009. Floral variation and floral genetics in basal angiosperms. *American Journal of Botany* 96: 110-128.

Stevens, P. F. (2001 onwards). Angiosperm phylogeny website. Version 12, July 2012 [and more or less continuously updated since]. http://www.mobot.org/MOBOT/research/APweb/.

Taberlet, P., L. Gielly, G. Pautou, and J. Bouvet. 1991. Universal primers for amplification of 3 noncoding regions of chloroplast DNA. *Plant Molecular Biology* 17: 1105-1109.

Techen, N., Z. Pan, B. E. Scheffier, and I. A. Khan. 2009. Detection of *Illicium anisatum* as adulterant of *Illicium verum*. *Planta Medica* 75: 392-395.

Ward, P., I. Small, J. Smith, P. Suter, and R. Dutkowski. 2005. Oseltamivir (Tamiflu (r)) and its potential for use in the event of an influenza pandemic. *Journal of Antimicrobial Chemotherapy* 55: 5-21.

Yandell, M., and D. Ence. 2012. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* 13: 329-342.

Zhang, J., M. Chen, X. Dong, R. Lin, J. Fan, and Z. Chen. 2015. Evaluation of four commonly used DNA barcoding loci for Chinese medicinal plants of the family Schisandraceae. *Plos One* 10: e0125574.

Zhang, Q., Y. Liu, and Sodmergen. 2003. Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. *Plant and Cell Physiology* 44: 941-951.

TABLES

Table 1. Plastomes used in *Illicium* and basal angiosperm comparative analyses. Plastomes with an accession number beginning with NC can be found in GenBank RefSeq. Plastomes with an accession number beginning with SRX can be found in GenBank SRA, where the raw sequence data has been deposited.

Order	**Species**	**Accession Number**	**Plastome size (bp)**
Amborellales	*Amborella trichopoda*	NC_005086	162,686
Nymphaeales	*Nymphaea alba*	NC_006050	159,930
	Nymphaea mexicana	NC_024542	159,962
	Nuphar advena	NC_008788	160,866
	Trithuria inconspicua	NC_020372	165,389
Austrobaileyales	*Illicium cubense*	SRX1317965	147,844
	Illicium floridanum	SRX1317966	148,097
	Illicium oligandrum	NC_009600	148,553
	Illicium ekmanii	SRX1317968	147,467
	Illicium henryi	SRX1317964	148,147

Table 2. Summary of plastome comparisons at each taxonomic level.

Old World *Illicium*	**New World *Illicium***	**All *Illicium***	**Basal Angiosperms**
Illicium oligandrum	*Illicium cubense*	*Illicium ekmanii*	*Amborella trichopoda*
Illicium henryi	*Illicium floridanum*	*Illicium cubense*	*Nymphaea alba*
	Illicium ekmanii	*Illicium floridanum*	*Nymphaea mexicana*
		Illicium oligandrum	*Nuphar advena*
		Illicium henryi	*Trithuria inconspicua*
			Illicium cubense
			Illicium floridanum
			Illicium oligandrum
			Illicium ekmanii
			Illicium henryi

Table 3. Analysis of *Illicium* plastome assemblies.

Species	% bp missing[1]	# reads mapped[2]	Avg. depth of coverage[3]	Max. depth of coverage[4]
Illicium cubense	0.31	58,876	44	167
Illicium floridanum	0.20	36,151	27	100
Illicium ekmanii	0.37	67,305	50	170
Illicium henryi	0.14	118,595	88	244

1. Percentage of base pairs missing from assembly, relative to reference sequence
2. Number of reads from raw data mapped to reference sequence during assembly. The number of reads mapped does not affect the size of the plastome, but indicates depth of coverage.
3. Average depth of reads mapped to reference sequence
4. Maximum depth of reads mapped to reference sequence

Table 4. Non-coding regions that were not sequenced via next-generation sequencing in each new *Illicium* plastome.

Region[2]	***I. cubense***[1]	***I. ekmanii***	***I. floridanum***	***I. henryi***
*ndhF-rpl32**	X	X		X
*ndhC-trnV**	X	X	X	
*rps16-trnQ**			X	
*trnT-psbD**	X		X	
atpF-atpH	X	X		
rps8-rpl14	X	X	X	
trnL-ndhF	X	X	X	X
trnS-trnG			X	
petN-psbM				X
trnT-trnL	X	X	X	X
ycf2-trnL	X	X	X	X

1. Regions that failed to sequence in a plastome are marked with an X. Regions that successfully sequenced are not marked.

2. Regions that were listed as top performers across all angiosperms in Shaw et al. (2014) are marked with an asterisk (*) and are listed in order of greatest to least variable.

Table 5. Primers designed for *Illicium* regions that failed to sequence during initial sequencing. These primers are in the process of being tested for utility.

Region	**Forward primer 5'-3'**	**Reverse primer 3'-5'**
atpF-atpH	CTCCTCCGCGTAGTTCTTCC	GCTTCCGTTATTGCTGCTGG
ndhC-trnV	CCTTCACGAATCGGGGCTAA	CCGAGAAGGTCTACGGTTCG
ndhF-rpl32	ACAAGCAGGAGTCCCAATCC	ACTGCGGTCCAATATCCCTT
petN-psbM	TTGCTTGGGCGGCTTTAATG	TGCTACTGCACTGTTCATTC
rps16-trnQ	CGCACGTTGCTTTCTACCAC	GTTCGAATCCTTCCGTCCCA
rps8-rpl14	AATTCGTAGACCGGGTCTGC	TGCGATCGCTCGGGAATTAA
trnS-trnG	CGCTTTAGTCCACTCAGCCA	TAGCTTGGAAGGCTAGGGGT
trnT-psbD	GCACGAAACGCCAGTCTTAG	AGGAACTGGCCAATCCATGG
trnT-trnL	ACCTCTGAGCTAAGCAGGCT	AGCGTCTACCAATTTCGCCA

Table 6. Top ten most potentially informative non-coding regions at different taxonomic levels in *Illicium.*

New World *Illicium*[2]	**Normalized PIC value[1]**	**Old World *Illicium*[3]**	**Normalized PIC value**	**All *Illicium*[4]**	**Normalized PIC value**
petN-psbM	9.48	*petN-psbM*	8.84	*petN-psbM*	10.83
rpl32-trnL[5]	8.23	*rps16-trnQ*[5, 6]	8.30	*rpl32-trnL*[5]	6.36
cemA-petA	7.23	*petA-psbJ*	6.14	*cemA-petA*	6.36
psbM-trnD	5.24	*trnS-trnG*	4.69	*petB* intron	3.52
trnM-atpE	4.49	*cemA-petA*	4.69	*psaC-ndhE*	3.38
trnQ-psbK	3.99	*petB* intron	4.15	*trnQ-psbK*	3.38
matK-rps16[6]	2.99	*trnT-psbD*	3.43	*psbM-trnD*	3.11
psbK-psbI	2.99	*ndhC-trnV*	3.25	*trnT-psbD*[5]	3.11
atpH-atpI[6]	2.74	*matK-rps16*[6]	2.89	*trnM-atpE*	2.98
trnH-psbA	2.74	*atpH-atpI*[6]	2.71	*matK-rps16*[6]	2.71

1. The normalized PIC value represents the percentage contribution of each NC-cpDNA region to the overall variability in a lineage (Shaw et al. 2014), and allows for comparisons between lineages.
2. Comparison between *I. ekmanii, I. cubense,* and *I. floridanum.*
3. Comparison between *I. henryi* and *I. oligandrum.*
4. Comparison between *I. henryi, I. oligandrum, I. ekmanii, I. cubense,* and *I. floridanum.*
5. *rpl32-trnL, rps16-trnQ,* and *trnT-psbD* were top performers across all angiosperms from the Shaw et al. 2014 analysis.
6. *matK-rps16 (trnK2-rps16* in Shaw et al. 2014), *atpH-atpI,* and *rps16-trnQ* were top performers from the Shaw et a. 2014 comparison between two basal angiosperm plastomes, *Nymphaea alba* and *Nuphar advena.*

Table 7. Top potentially informative regions in all basal angiosperm plastomes surveyed, and regions excluded.

Basal Angiosperms[1]	**Normalized PIC value[2]**	**Regions Excluded[3]**
psbE-petL[5]	4.46	*atpH-atpI*[6]
rpoB-trnC[5]	3.92	*matK-rps16*
matK	3.76	*ndhA intron*
trnE-trnT	3.25	*petD-rpoA*
psbM-trnD	3.09	*petN-psbM*
trnC-petN	3.06	*rpl32-trnL*[6]
rpl16 intron[4]	2.71	
ycf3-trnS	2.70	
trnF-ndhJ	2.57	
accd-psaI[5]	2.56	

1. Most variable regions in all basal angiosperm plastomes available
2. The normalized PIC value represents the percentage contribution of each NC-cpDNA region to the overall variability in a lineage (Shaw et al. 2014), and allows for comparisons between lineages.
3. Regions excluded from the basal angiosperm comparison analysis due to inability to confidently align
4. *rpl16 intron* was a top performer across all angiosperms from the Shaw et al. 2014 analysis.
5. *psbE-petL, rpoB-trnC,* and *accD-psaI* were top performers from the Shaw et a. 2014 comparison between two basal angiosperm plastomes, *Nymphaea alba* and *Nuphar advena.*
6. *atpH-atpI* and *rpl32-trnL,* excluded here, were top performers in the Shaw et al. 2014 analysis.

FIGURES

Figure 1. Gene order and content in the *Illicium oligandrum* plastome, obtained from GenBank (NC_009600). *I. oligandrum* was used as the reference plastome for assembly of newly sequenced *I. cubense, I. ekmanii, I. floridanum,* and *I. henryi* plastomes.

Figure 2. Flower morphological variation in New and Old World *Illicium.* Top row: representatives of New World *Illicium.* Bottom row: representatives of Old World *Illicium.* These specimens demonstrate the varied flower morphology present in both the Old and New World clades.

Photo credits, top row, left to right: A. B. Morris, J. Ruter, R. Abbott

Photo credits, bottom row, left to right: R. Pooma, Jade Lau, FRIM Malaysia, Colesville Nursery, VA

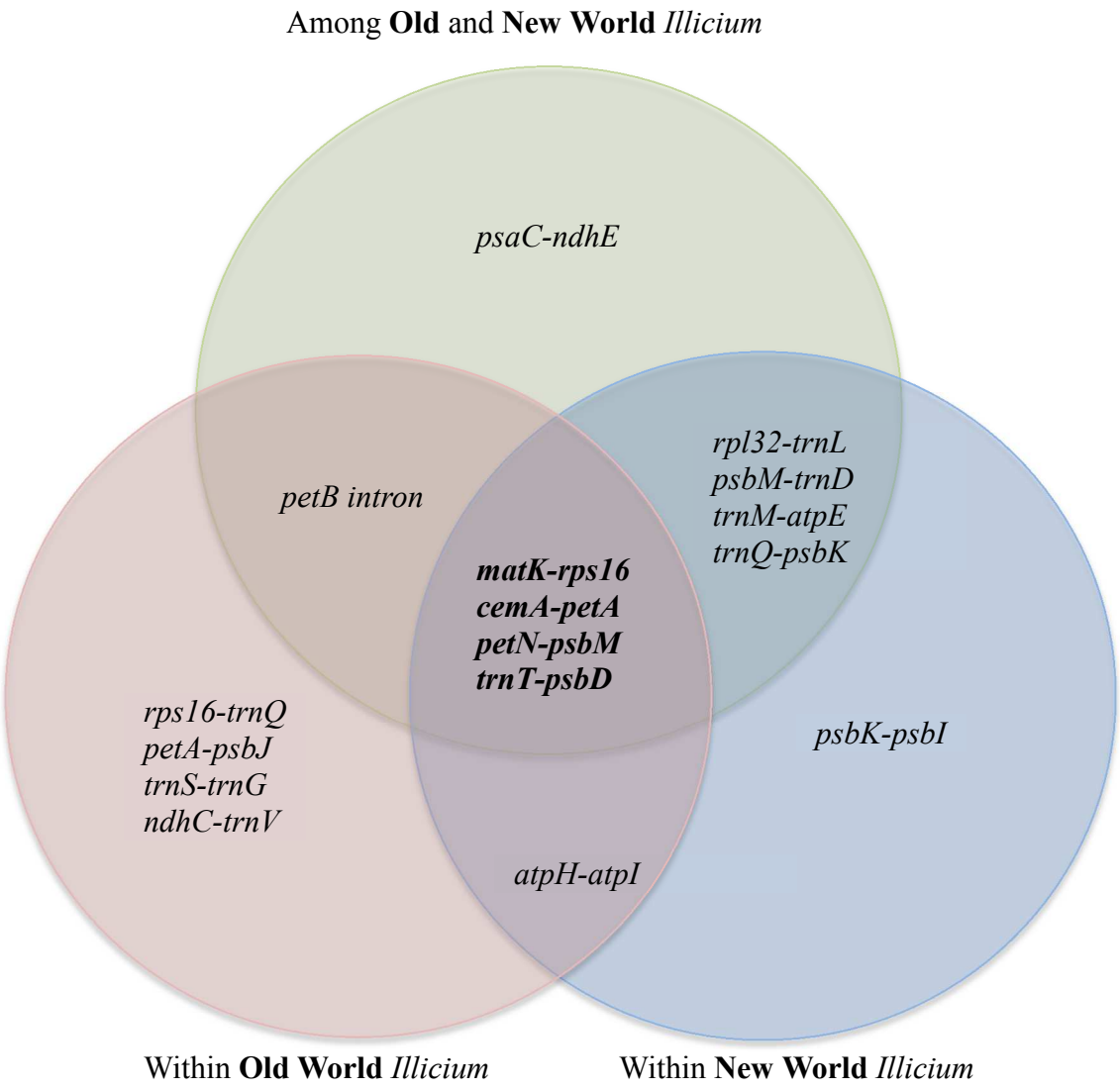

Figure 3. Comparison of the top variable regions in each *Illicium* analysis, sorted by normalized PIC value, found in Table 6. The regions variable across all levels of taxonomic analysis in *Illicium* are: *matK-rps16, cemA-petA, petN-psbM,* and *trnT-psbD.* This comparison demonstrates the difference in variability between groups at low taxonomic levels.

CHAPTER 3: PIC COUNTER: A SIMPLE PROGRAM TO COUNT POLYMORPHISMS IN DNA ALIGNMENTS FOR COMPARATIVE PLASTOMICS ANALYSES

Abstract

- *Premise of study:* As next-generation sequencing becomes more accessible to researchers, it becomes more important that researchers have the appropriate tools for a given study. Comparative plastomics have become a common tool for screening plastomes for potentially informative regions.
- *Methods and results:* The PIC Counter scripts were developed using Perl and are executable on the command line. The PIC Counter scripts count number of SNPs in a pairwise or multiple alignment, length of the alignment, and position and base pairs of SNPs. The script for the pairwise alignment also counts indels.
- *Conclusions:* The PIC Counter scripts make the plastome screening process easier and faster, and decrease the likelihood of human error in counts. The scripts are not interactive, require no additional downloaded software, minimal computational knowledge, and little computer RAM.

Introduction

Plant DNA has been widely used by plant systematists for almost two decades (Straub et al., 2012). During most of that time, Sanger sequencing was the most common method of obtaining DNA sequences. In recent years, next-generation sequencing (NGS), also known as high-throughput sequencing, has become more commonplace and more accessible to researchers (Godden et al., 2012). NGS is a powerful tool with the potential to revolutionize plant systematics (Soltis et al., 2013). However, challenges still exist. Limitations such as accessibility of computational support and technical expertise can limit the availability of NGS for researchers lacking the resources and time. NGS sequence data also requires computational power and bioinformatics experience, and the learning curve is steep. Therefore, many labs are using a screening process in which two to three plastomes are sequenced via NGS and visually screened and counted SNPs (single nucleotide polymorphisms) and indels, in order to determine which regions will be most phylogenetically useful for a particular group (Sarkinen and George, 2013, Shaw et al., 2014).

Plastome comparisons involve extracting DNA regions of interest from all taxa to be analyzed, then aligning those regions and scoring the alignments for potentially informative characters (PICs). Manually scoring a multiple alignment for SNPs can be time consuming and taxing on the researcher. Variant calling programs are available, but require downloading, setup, or program-specific file formatting, and the time spent on setup is worth it only if there are many, perhaps dozens, of plastomes to be analyzed (see

Carbonell-Caballero et al., 2015, for an example). In many cases, researchers are assessing a few plastomes for SNPs and indels, and therefore, the time commitment for many software packages may not be justified.

The PIC Counter programs are simple Perl scripts that make the screening process faster and easier. The PIC Counter 2X is a program that scores a pairwise alignment for SNPs and indels. The multiple alignment PIC Counter scripts score multiple alignments with three or more DNA sequences and counts SNPs present in the alignment, though for multiple alignments, the researcher will have to count indels by eye, an easy task compared to counting SNPs. The multiple alignment PIC Counter is easily edited to assess multiple alignments from three taxa and up, and adding the ability of the script to count indels in multiple alignments will negate the ability of the researcher to easily and quickly edit the script to accommodate any number of sequences in the alignment. Two examples of the multiple alignment PIC Counter are included: PIC Counter 3x, which parses multiple alignments involving three taxa, and PIC counter 10x, which parses multiple alignments involving ten taxa. There is no limit to how many taxa the multiple alignment PIC Counter will analyze, except perhaps the availability of RAM on the computer used to run the script.

Methods and Materials

PIC Counter can be used on any multiple alignment in any form of genetic code, including DNA and RNA. It is run from the command line and yields results quickly, even on large alignments with many taxa. PIC Counter consists of Perl scripts, and all scripts and example files are contained in a public GitHub repository (https://github.com/rayneleonard/Counting_PICs) under a BSD open-source license.

The PIC Counter scripts were developed using Perl version 5.18.2 on Mac OS X. The Perl executables are PIC_counter_2x.pl, PIC_counter_3x.pl, and PIC_counter_10x.pl. The input file required is a multiple alignment in FASTA format, and the file name should be entered on the command line when executing the Perl script. An output file, if not otherwise specified in the script, will be saved under the file name output_2x.txt, output_3x.txt, or output_10x.txt. The output file will include total number of SNPs, total number of indels (for the pairwise alignment script), position and base pairs of SNPs, and length of the alignment.

The PIC Counter scripts are easily edited to accommodate any number of taxa in the multiple alignment. The 3X and 10X programs are very similar, but were both included to demonstrate how the programs can be quickly edited via copying and pasting. The pairwise alignment script includes number of indels as well as substitutions in the alignment; the multiple alignment scripts do not count indels in order to preserve the simplicity of the programs. However, indels in multiple alignments are easily counted manually. To validate the accuracy of the PIC Counter scripts, sample input and output

files were included. The sample alignments are short to allow for easy manual double-checking.

Conclusions

As NGS becomes more accessible for researchers, more plastomes will be publically available in online databases, and a plastome screening process will remain a valuable first step for researchers selecting markers for phylogenetic and phylogeographic analyses. It is important that researchers have reliable tools for screening analyses, and these scripts contribute to the available options for researchers. These scripts are a good choice for researchers seeking quick comparative data, with little time commitment for setup and file formatting, and little required RAM.

The PIC Counter programs make the plastome screening process easier by eliminating manually counting SNPs in a multiple alignment. These scripts are not interactive, do not rely on internet access, and require no additional downloaded software or computational knowledge. The scripts facilitate rapid assessment of potentially informative regions while decreasing the likelihood of human error in counts.

REFERENCES

Carbonell-Caballero, J., R. Alonso, V. Ibanez, J. Terol, M. Talon, and J. Dopazo. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus *Citrus. Molecular Biology and Evolution* 32:2015-2035.

Godden, G. T., I. E. Jordon-Thaden, S. Chamala, A. A. Crowl, N. Garcia, C. C. Germain-aubrey, J. M. Heaney, et al. 2012. Making next-generation sequencing work for you: approaches and practical considerations for marker development and phylogenetics. *Plant Ecology & Diversity* 5: 427-450.

Saerkinen, T., and M. George. 2013. Plastid marker variation: can complete plastid genomes from closely related species help? *Plos One* 8: e82266.

Shaw, J., H. L. Shafer, O. R. Leonard, M. J. Kovach, M. Schorr, and A. B. Morris. 2014. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. *American Journal of Botany* 101: 1987-2004.

Soltis, D. E., M. A. Gitzendanner, G. Stull, M. Chester, A. Chanderbali, S. Chamala, I. Jordon-Thaden, et al. 2013. The potential of genomics in plant systematics. *Taxon* 62: 886-898.

Straub, S. C. K., M. Parks, K. Weitemier, M. Fishbein, R. C. Cronn, and A. Liston. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349-364.

CHAPTER 4: OVERALL CONCLUSIONS

A comparison of whole chloroplast genomes (i.e., plastomes) across different taxonomic levels in *Illicium* and the basal angiosperms was completed. Four *Illicium* plastomes were sequenced, and six plastomes, representative of all orders of the basal angiosperms, were attained from GenBank. The most variable potentially informative regions for phylogenetic analyses differed across taxonomic levels in *Illicium,* as well as across the basal angiosperms. Perl programs were developed to make the plastome screening process quicker and with less human error. Screening of a relatively small number of plastomes for variable regions specific to a group of organisms will continue to be fruitful, as researcher continue to choose Sanger sequencing for research. There is limited recent literature specific to phylogenetic relationships within the basal angiosperms, and this study contributes significantly to the literature by detailing the most potentially informative non-coding chloroplast DNA regions for future phylogenetic analyses in *Illicium* and the basal angiosperms via comparative plastomics. By providing *Illicium* plastomes, the amount of data for basal angiosperm specific research is increased.

APPENDICES

APPENDIX A: Compiled data in tabular form.

Old World *Illicium (I. henryi, I. oligandrum)* comparative analysis data. Total number of PICs: 554.

Region	Total PICs	Length	%var	Norm value
petN-psbM	49	1081	0.0453284	8.84476534
rps16-trnQ	46	1711	0.02688486	8.3032491
petA-psbJ	34	1192	0.02852349	6.13718412
trnS-trnG	26	964	0.02697095	4.69314079
cemA-petA	26	225	0.11555556	4.69314079
petB intron	23	787	0.0292249	4.15162455
trnT-psbD	19	1183	0.01606086	3.42960289
ndhC-trnV	18	1396	0.01289398	3.24909747
matK-rps16	16	2259	0.00708278	2.88808664
atpH-atpI	15	1049	0.01429933	2.70758123
trnM-atpE	14	238	0.05882353	2.52707581
rpl32-trnL	14	1805	0.00775623	2.52707581
trnQ-psbK	14	339	0.04129794	2.52707581
rps15-ycf1	13	406	0.0320197	2.3465704
psbK-psbI	11	393	0.02798982	1.98555957
rpoB-trnC	10	1331	0.00751315	1.80505415
atpF intron	10	823	0.01215067	1.80505415
psbE-petL	9	1190	0.00756303	1.62454874
trnH-psbA	9	387	0.02325581	1.62454874
atpB-rbcL	8	758	0.01055409	1.44404332
trnF-ndhJ	8	656	0.01219512	1.44404332
accD-psaI	7	646	0.01083591	1.26353791
rps16 intron	7	853	0.00820633	1.26353791
clpP intron *I*	7	796	0.00879397	1.26353791
ndhA intron	7	1057	0.00662252	1.26353791

Old World comparative analysis data, continued.

Region	Total PICs	Length	%var	Norm value
trnC-petN	7	899	0.00778643	1.26353791
ycf3-trnS	7	766	0.00913838	1.26353791
rpl16 intron	6	1008	0.00595238	1.08303249
clpP intron *II*	6	649	0.00924499	1.08303249
trnE-trnT	6	660	0.00909091	1.08303249
trnS-psbZ	5	337	0.0148368	0.90252708
ycf3 intron *II*	5	767	0.0065189	0.90252708
atpF-atpH	5	536	0.00932836	0.90252708
psbM-trnD	4	1128	0.0035461	0.72202166
trnS-rps4	4	230	0.0173913	0.72202166
psaA-ycf3	4	619	0.00646204	0.72202166
psbA-matK	4	585	0.00683761	0.72202166
psbZ-trnG	4	259	0.01544402	0.72202166
trnL intron	4	519	0.00770713	0.72202166
petG-trnW	3	123	0.02439024	0.54151625
rps8-rpl14	3	226	0.01327434	0.54151625
trnG intron	3	732	0.00409836	0.54151625
ycf3 intron *I*	3	745	0.00402685	0.54151625
petL-petG	3	183	0.01639344	0.54151625
rpl16-rps3	3	166	0.01807229	0.54151625
rpoC1 intron	3	737	0.00407056	0.54151625
rps2-rpoC2	3	204	0.01470588	0.54151625
trnD-trnY	2	373	0.00536193	0.36101083
psbC-trnS	2	234	0.00854701	0.36101083
ndhD-psaC	2	122	0.01639344	0.36101083
ndhG-ndhI	2	368	0.00543478	0.36101083
psaJ-rpl33	2	436	0.00458716	0.36101083
rbcL-accD	2	640	0.003125	0.36101083

Old World comparative analysis data, continued.

Region	Total PICs	Length	%var	Norm value
rpl20-rps12	2	760	0.00263158	0.36101083
rpl36-infA	2	119	0.01680672	0.36101083
trnV intron	2	606	0.00330033	0.36101083
trnW-trnP	2	172	0.01162791	0.36101083
psbB-psbT	1	192	0.00520833	0.18050542
rps18-rpl20	1	215	0.00465116	0.18050542
trnG-trnR	1	155	0.00645161	0.18050542
trnL-trnF	1	259	0.003861	0.18050542
trnV-trnM	1	208	0.00480769	0.18050542
ccsA-ndhD	1	294	0.00340136	0.18050542
clpP-psbB	1	433	0.00230947	0.18050542
ndhE-ndhG	1	264	0.00378788	0.18050542
ndhI-ndhA	1	79	0.01265823	0.18050542
ndhJ-ndhK	1	107	0.00934579	0.18050542
petB-petD	1	164	0.00609756	0.18050542
psaI-ycf4	1	412	0.00242718	0.18050542
psbJ-psbL	1	119	0.00840336	0.18050542
rpoC2-rpoC1	1	156	0.00641026	0.18050542
rps14-psaB	1	143	0.00699301	0.18050542
trnfM-rps14	1	159	0.00628931	0.18050542
trnG-trnfM	1	182	0.00549451	0.18050542
trnL-ccsA	1	102	0.00980392	0.18050542
trnP-psaJ	1	339	0.00294985	0.18050542
rpl14-rpl16	0	147	0	0
atpI-rps2	0	198	0	0
infA-rps8	0	121	0	0

Old World comparative analysis data, continued.

Region	Total PICs	Length	%var	Norm value
ndhH-rps15	0	104	0	0
petD intron	0	729	0	0
petD-rpoA	0	169	0	0
psbN-psbH	0	104	0	0
rpl33-rps18	0	106	0	0
rps11-rpl36	0	116	0	0
rps12-clpP	0	152	0	0
rps4-trnT	0	381	0	0
trnR-atpA	0	123	0	0
ycf4-cemA	0	900	0	0

New World *Illicium (I. cubense, I. ekmanii, I. floridanum)* comparative analysis data. Total number of PICs: 401.

Region	Total PICs	Length	%var	Norm value
petN-psbM	38	1028	0.03696498	9.47630923
rpl32-trnL	33	1558	0.021181	8.22942643
cemA-petA	29	223	0.13004484	7.2319202
psbM-trnD	21	1128	0.01861702	5.23690773
trnM-atpE	18	262	0.06870229	4.48877805
trnQ-psbK	16	338	0.04733728	3.99002494
matK-rps16	12	1598	0.00750939	2.9925187
psbK-psbI	12	405	0.02962963	2.9925187
atpH-atpI	11	1039	0.0105871	2.74314214
trnH-psbA	11	376	0.02925532	2.74314214
trnE-trnT	10	662	0.01510574	2.49376559
psbA-matK	10	580	0.01724138	2.49376559
psbE-petL	9	1210	0.00743802	2.24438903
rps16 intron	8	151	0.05298013	1.99501247
rpoB-trnC	8	1291	0.00619675	1.99501247
ndhA intron	8	1060	0.00754717	1.99501247
rpl16 intron	7	1027	0.00681597	1.74563591
ycf3 intron *II*	7	773	0.00905563	1.74563591
ycf3-trnS	7	779	0.00898588	1.74563591
psaJ-rpl33	6	437	0.01372998	1.49625935
trnF-ndhJ	5	661	0.0075643	1.24688279
ycf3 intron I	5	745	0.00671141	1.24688279
trnC-petN	5	895	0.00558659	1.24688279
trnfM-rps14	5	157	0.03184713	1.24688279
clpP intron *II*	4	657	0.00608828	0.99750623
atpF intron	4	837	0.00477897	0.99750623
ndhG-ndhI	4	370	0.01081081	0.99750623

New World comparative analysis data, continued.

Region	Total PICs	Length	%var	Norm value
petB intron	4	777	0.00514801	0.99750623
rpl16-rps3	4	161	0.02484472	0.99750623
rpl20-rps12	4	762	0.00524934	0.99750623
trnW-trnP	4	172	0.02325581	0.99750623
atpB-rbcL	3	757	0.00396301	0.74812968
trnD-trnY	3	378	0.00793651	0.74812968
accD-psaI	3	636	0.00471698	0.74812968
psaC-ndhE	3	289	0.01038062	0.74812968
trnG-trnR	3	155	0.01935484	0.74812968
trnL-ccsA	3	102	0.02941176	0.74812968
petD intron	3	726	0.00413223	0.74812968
rpoC1 intron	3	725	0.00413793	0.74812968
rps18-rpl20	3	225	0.01333333	0.74812968
trnG intron	2	732	0.00273224	0.49875312
clpP intron *I*	2	800	0.0025	0.49875312
petG-trnW	2	118	0.01694915	0.49875312
psbB-psbT	2	192	0.01041667	0.49875312
rps14-psaB	2	156	0.01282051	0.49875312
rps15-ycf1	2	408	0.00490196	0.49875312
atpI-rps2	2	199	0.01005025	0.49875312
ndhE-ndhG	2	265	0.00754717	0.49875312
ndhI-ndhA	2	79	0.02531646	0.49875312
psaI-ycf4	2	414	0.00483092	0.49875312
psbZ-trnG	2	257	0.0077821	0.49875312
rbcL-accD	2	640	0.003125	0.49875312
rps2-rpoC2	2	202	0.00990099	0.49875312
rps4-trnT	2	382	0.0052356	0.49875312

New World comparative analysis data, continued.

Region	Total PICs	Length	% var	Norm value
trnL intron	2	518	0.003861	0.49875312
trnR-atpA	2	123	0.01626016	0.49875312
trnG-trnfM	1	181	0.00552486	0.24937656
trnS-rps4	1	230	0.00434783	0.24937656
clpP-psbB	1	432	0.00231481	0.24937656
infA-rps8	1	122	0.00819672	0.24937656
petB-petD	1	170	0.00588235	0.24937656
psaA-ycf3	1	619	0.00161551	0.24937656
rpl36-infA	1	119	0.00840336	0.24937656
rpoC2-rpoC1	1	154	0.00649351	0.24937656
rps12-clpP	1	157	0.00636943	0.24937656
trnL-trnF	1	255	0.00392157	0.24937656
trnP-psaJ	1	338	0.00295858	0.24937656
trnS-psbZ	1	328	0.00304878	0.24937656
trnV-trnM	1	208	0.00480769	0.24937656
rpl14-rpl16	0	147	0	0
trnV intron	0	209	0	0
ccsA-ndhD	0	295	0	0
ndhD-psaC	0	123	0	0
ndhH-rps15	0	105	0	0
ndhJ-ndhK	0	106	0	0
petD-rpoA	0	169	0	0
petL-petG	0	180	0	0
psbC-trnS	0	233	0	0
psbJ-psbL	0	121	0	0
psbN-psbH	0	104	0	0
rpl33-rps18	0	116	0	0
rps11-rpl36	0	116	0	0

All *Illicium (I. cubense, I. ekmanii, I. floridanum, I. henryi, I. oligandrum)* comparative analysis data. Number of PICs: 739.

Region	Total PICs	Length	%var	Norm value
petN-psbM	80	1093	0.07319305	10.8254398
rpl32-trnL	47	1817	0.02586681	6.35994587
cemA-petA	47	228	0.20614035	6.35994587
petB intron	26	787	0.03303685	3.51826793
psaC-ndhE	25	298	0.08389262	3.38294993
trnQ-psbK	25	346	0.07225434	3.38294993
psbM-trnD	23	1129	0.02037201	3.11231394
trnT-psbD	23	1182	0.01945854	3.11231394
trnM-atpE	22	262	0.08396947	2.97699594
matK-rps16	20	2278	0.00877963	2.70635995
rps8-rpl14	17	227	0.07488987	2.30040595
trnH-psbA	17	387	0.04392765	2.30040595
psbE-petL	16	1209	0.01323408	2.16508796
atpF-atpH	16	549	0.0291439	2.16508796
atpH-atpI	15	1057	0.01419111	2.02976996
rps16 intron	14	851	0.01645123	1.89445196
atpF intron	13	842	0.01543943	1.75913396
psbK-psbI	13	408	0.03186275	1.75913396
rps15-ycf1	13	408	0.03186275	1.75913396
trnE-trnT	13	662	0.01963746	1.75913396
psbA-matK	13	585	0.02222222	1.75913396
ycf3-trnS	13	779	0.01668806	1.75913396
rpoB-trnC	12	1331	0.00901578	1.62381597
trnF-ndhJ	11	662	0.01661631	1.48849797
ndhA intron	11	1060	0.01037736	1.48849797
rpl16 intron	10	1027	0.0097371	1.35317997

All *Illicium* comparative analysis data, continued.

Region	Length	Total PIC	norm value	Region
psaJ-rpl33	9	436	0.0206422	1.21786198
trnC-petN	9	900	0.01	1.21786198
ycf3 intron *II*	8	773	0.01034929	1.08254398
accD-psaI	7	645	0.01085271	0.94722598
ycf3 intron *I*	7	745	0.00939597	0.94722598
clpP intron *I*	7	806	0.00868486	0.94722598
atpB-rbcL	6	763	0.0078637	0.81190798
clpP intron *II*	6	826	0.00726392	0.81190798
ndhG-ndhI	6	369	0.01626016	0.81190798
trnS-psbZ	6	337	0.01780415	0.81190798
rpl16-rps3	6	166	0.03614458	0.81190798
trnD-trnY	5	382	0.01308901	0.67658999
psaA-ycf3	5	619	0.00807754	0.67658999
rpl20-rps12	5	765	0.00653595	0.67658999
rpoC1 intron	5	735	0.00680272	0.67658999
trnfM-rps14	5	158	0.03164557	0.67658999
trnL intron	5	519	0.00963391	0.67658999
trnG intron	4	733	0.00545703	0.54127199
trnG-trnR	4	157	0.02547771	0.54127199
psbZ-trnG	4	259	0.01544402	0.54127199
trnL-ccsA	4	102	0.03921569	0.54127199
trnW-trnP	4	173	0.02312139	0.54127199
rps18-rpl20	3	225	0.01333333	0.40595399
trnS-rps4	3	232	0.01293103	0.40595399
petG-trnW	3	122	0.02459016	0.40595399
rps2-rpoC2	3	203	0.01477833	0.40595399
psaI-ycf4	3	414	0.00724638	0.40595399
rbcL-accD	3	640	0.0046875	0.40595399

All *Illicium* comparative analysis data, continued.

Region	Length	Total PIC	norm value	Region
psbC-trnS	2	234	0.00854701	0.27063599
rps14-psaB	2	156	0.01282051	0.27063599
trnG-trnfM	2	182	0.01098901	0.27063599
trnL-trnF	2	260	0.00769231	0.27063599
atpI-rps2	2	198	0.01010101	0.27063599
ndhD-psaC	2	122	0.01639344	0.27063599
ndhE-ndhG	2	165	0.01212121	0.27063599
petD intron	2	725	0.00275862	0.27063599
petB-petD	2	170	0.01176471	0.27063599
petL-petG	2	182	0.01098901	0.27063599
rps4-trnT	2	381	0.00524934	0.27063599
trnP-psaJ	2	339	0.00589971	0.27063599
trnR-atpA	2	123	0.01626016	0.27063599
trnV-trnM	2	208	0.00961538	0.27063599
psbB-psbT	1	194	0.00515464	0.135318
rpl36-infA	1	119	0.00840336	0.135318
trnV intron	1	607	0.00164745	0.135318
ccsA-ndhD	1	294	0.00340136	0.135318
clpP-psbB	1	433	0.00230947	0.135318
infA-rps8	1	121	0.00826446	0.135318
ndhI-ndhA	1	79	0.01265823	0.135318
ndhJ-ndhK	1	107	0.00934579	0.135318
psbJ-psbL	1	121	0.00826446	0.135318
rpoC2-rpoC1	1	155	0.00645161	0.135318
rps12-clpP	1	157	0.00636943	0.135318
rpl14-rpl16	0	149	0	0
ndhH-rps15	0	104	0	0
petD-rpoA	0	169	0	0

All *Illicium* comparative analysis data, continued.

Region	Length	Total PIC	norm value	Region
psbN-psbH	0	104	0	0
rpl33-rps18	0	106	0	0
rps11-rpl36	0	106	0	0
ycf4-cemA	0	900	0	0

Basal angiosperm *(I. cubense, I. ekmanii, I. floridanum, I. henryi, I. oligandrum, N. advena, N. alba, N. mexicana, T. inconspicua, A. trichopoda)* comparative analysis data. Six regions were excluded from analysis and not included here, due to inability to confidently align: *atpH-atpI, matK-rps16, ndhA intron, petD-rpoA, petN-psbM, rpl32-trnL.* Barcodes *matK* and *rbcL* are included for variability comparison. Total number of PICs: 12,934.

Region	Length	Total PICs	norm value
petN-psbM	1769	1518	11.736508
matK-rps16	3457	751	5.8064017
psbE-petL	1563	577	4.46111025
atpH-atpI	1896	522	4.0358744
rpoB-trnC	1736	507	3.91990104
matK	1551	486	3.75753827
trnE-trnT	1096	420	3.2472553
psbM-trnD	1223	400	3.09262409
trnC-petN	1227	396	3.06169785
ndhA intron	1270	364	2.8145879
rpl16 intron	1261	350	2.70604608
ycf3-trnS	1093	349	2.69831452
trnF-ndhJ	819	333	2.57460956
accD-psaI	932	331	2.55914644
atpB-rbcL	877	326	2.52048863
rbcL-accD	851	325	2.51275707
petB intron	930	279	2.1571053
rps16 intron	970	273	2.11071594
psbA-matK	794	261	2.01793722
atpF intron	1477	256	1.97927942
rpl20-rps12	948	253	1.95608474
psaA-ycf3	876	236	1.82464821
trnG intron	900	232	1.79372197
trnH-psbA	715	231	1.78599041
rps15-ycf1	502	230	1.77825885

Basal angiosperm comparative analysis data, continued.

Region	Length	Total PICs	norm value
psaC-ndhE	748	221	1.70867481
rpoC1 intron	854	221	1.70867481
ccsA-ndhD	619	218	1.68548013
psaI-ycf4	549	209	1.61589609
petD intron	810	208	1.60816453
ycf3 intron *II*	949	205	1.58496985
rbcL	1428	203	1.56950673
trnD-trnY	733	187	1.44580176
ycf3 intron *I*	852	185	1.43033864
ndhG-ndhI	434	184	1.42260708
psaJ-rpl33	508	175	1.35302304
trnS-psbZ	410	171	1.3220968
psbK-psbI	502	168	1.29890212
ndhE-ndhG	368	162	1.25251276
trnQ-psbK	412	152	1.17519715
trnL intron	682	146	1.12880779
trnL-trnF	510	146	1.12880779
trnV intron	655	146	1.12880779
trnP-psaJ	474	140	1.08241843
rps4-trnT	482	135	1.04376063
trnS-rps4	648	127	0.98190815
trnG-trnR	366	112	0.86593475
rps2-rpoC2	654	109	0.84274006
trnG-trnfM	238	109	0.84274006
psbC-trnS	271	105	0.81181382
psbZ-trnG	451	105	0.81181382
cemA-petA	397	102	0.78861914
rps18-rpl20	346	101	0.78088758

Basal angiosperm comparative analysis data, continued.

Region	Length	Total PICs	norm value
trnW-trnP	182	92	0.71130354
atpI-rps2	298	86	0.66491418
rpl14-rpl16	236	84	0.64945106
trnL-ccsA	157	84	0.64945106
petL-petG	214	80	0.61852482
rpl16-rps3	188	79	0.61079326
trnM-atpE	256	75	0.57986702
trnfM-rps14	174	72	0.55667234
trnV-trnM	230	67	0.51801454
petG-trnW	139	63	0.48708829
psbB-psbT	206	61	0.47162517
rps14-psaB	400	60	0.46389361
rpl33-rps18	156	58	0.44843049
infA-rps8	136	57	0.44069893
trnR-atpA	333	54	0.41750425
petB-petD	221	51	0.39430957
rps11-rpl36	148	51	0.39430957
rpoC2-rpoC1	171	49	0.37884645
rpl36-infA	146	44	0.34018865
ndhD-psaC	144	41	0.31699397
ndhH-rps15	127	40	0.30926241
ndhJ-ndhK	115	31	0.23967837
psbJ-psbL	132	27	0.20875213
psbN-psbH	109	25	0.19328901

APPENDIX B: Compiled data from literature review.

Author	Year	Family	Markers used	Length	PICs	Sequencing
Biswal	2012	Nymphaeaceae	matK	1524	Not	n/a
		Cabombaceae	ITS2	243	reported	
Borsch	2014	Nymphaeaceae	ITS	Not reported	238	Sanger
			Rps4-trnT-trnF		38	
Fan	2011	Schisandraceae	ITS	5829 total	258 total	Sanger
			matK			
			psbA-trnH			
			rbcL			
			rpl16			
			trnL-trnF			
Goremykin	2012	Several	61 coding, not listed	NR	NR	NGS
Iles	2012	Hydatellaceae	atpB	4122	NR	Sanger
			matK			
			ndhF			
			rbcL			
			ITS	717		
Iles	2014	Hydatellaceae	13 coding	Data from Iles 2012		n/a
Kim, C	2012	Cabombaceae	ITS	652	2	Sanger
			trnT-trnF	1417	26	
Kim, J	2012	Schisandraceae	rbcL	790	10	Sanger
			ITS	763	35	
Maia	2014	429 taxa	rDNA	NR	NR	n/a
Meizi	2012	Schisandraceae	matK	NR	NR	Sanger
			rbcL			
			psbA-trnH			
			ITS			
Moore	2010	86 taxa	83 coding	66741 total	NR	NGS
Moore	2011	244 taxa	IR region	25k	NR	NGS
Morton	2011	247 genera	Xdh, nuclear	1265	1187	Sanger
Qiu	2010	376 genera	Mito genes	NR	NR	Sanger
Soltis	2011	330 families	17 coding, nr, mito, and plastid	25260 total	NR	Sanger
Zhang	2015	Schisandraceae	ITS	1223	369	Sanger
			trnH-psbA	579	94	
			matK	826	65	
			rbcL	672	29	

APPENDIX C: PIC Counter programs.

PIC Counter 2x script

```
#!/usr/bin/perl
#perl version 5.18.2


#writing a program that will count subs and indels in a fasta or txt
alignment file
#utilizes pairwise alignment

#printing length of indel to output file will be included in an update

#open gene input file
open(INPUT, "$ARGV[0]") || die "Can't find fasta file, try again; $!n";
#can make this go through all files in a directory
open (OUTPUT, '>output_2x.txt') || die "can't open output;$!n";

print OUTPUT "Type\t\tbp\t\tposition\n";

#put into array;
@input = <INPUT>;


#assigning vars
$refseqnt = 0;
$counter = 0;
$queryseqnt = 0;
$insert_length = 0;
$delete_length = 0;


$refseq_name = $input[0];
$refseq = $input[1];
$query_name = $input[2];
$queryseq = $input[3];

#chomp and clear out invisible characters here
chomp $refseq_name;
chomp $refseq;
chomp $query_name;
chomp $queryseq;


#1 and 3 need to be vars to start with then split on '', not in loop.
#this explodes the strings. Already single strings.

@refseq = split('', $refseq);
@queryseq = split ('', $queryseq);
#
# print "@refseq\n\n\n@queryseq";
$lengtharray = scalar @refseq;
```

```
while ($counter <= $lengtharray)
{
        $refseqnt = $refseq[$counter];
        #counter is how many things in arrays
        $queryseqnt = $queryseq[$counter];


        if ($refseqnt ne $queryseqnt)
        {
                if ($refseqnt eq "-")
                {
#               $prior_insert = $refseqnt;
                        if ($prior_insert eq "-")
                        {
                        ++$insert_length;
                        #print "Hello";
                        }
#               ++$insert_length;

                        else #prior_insert !eq -
                        {
                        $insert_location= $counter - $insert_length;
                        print OUTPUT "Insertion\t\t\t$insert_location\n";
                        $insert_length = 0;
                        ++$count_up_insertion;

                        }
                }

                if ($queryseqnt eq "-")
                {
                        if ($prior_delete eq "-")
                        {
                        ++$delete_length;
                        #print "Hello again";
                        }
                        else
                        {
                        $delete_location = $counter - $delete_length;
                        print OUTPUT "Deletion\t\t\t$delete_location\n";
                        $delete_length = 0;
                        ++$count_up_deletion;
                        }
                }

                if ($refseqnt ne "-" && $queryseqnt ne "-")
                {
                ++$count_up_snp;
                print OUTPUT
"SNP\t\t\t$refseqnt\t$queryseqnt\t$counter\n";
                }
        }
        $prior_insert = $refseqnt;
```

```
        $prior_delete = $queryseqnt;
        ++$counter;
}
$indels = $count_up_insertion + $count_up_deletion;
$snp = $count_up_snp;

print "Total number of indels for $refseq_name is $indels and total
number of SNPs is $snp\n\n";
print "Total length of array is $lengtharray\n";

#example file 2x_example.fasta length 63 bp
#indels 1
#SNPs 3
```

Example PIC Counter 2x input file:

```
>spec_1
CAACAAGTATTTAGTTCATCGGAATCGAAATAACAAGAATGGGGGTTTCTTTTCTCACATAAG
>spec_2
GAACAAGTATATAGTTCA---GAATCGAAATAACAAGAAGGGGGGTTTCTTTTCTCACATAAA
```

Example PIC Counter 2x output file:

```
Type            bp              position
SNP                     C       G       0
SNP                     T       A       10
Deletion                                18
SNP                     T       G       39
SNP                     G       A       62
```

PIC Counter 3x script

```
#!/usr/bin/perl
#using perl version 5.18.2

#writing a program that will count subs in a fasta or txt alignment
file

#in this program, I will treat each input[1], input[3], and so on
#and put them into an array together to be parsed.

#this program will use a multiple alignment with 3 sequences aligned in
fasta format
#the first sequence is used as the reference
#can be easily edited for more sequences in alignment; see 10x program

#note that this only counts SNPs, not insertions/deletions.

#open gene input file; needs to be in FASTA format
#pull in from command line
open(INPUT, "$ARGV[0]") || die "Can't find input file, try again; $!n";
#open output
open (OUTPUT, '>output_3x.txt') || die "can't open output;$!n";

print OUTPUT "\t\t\tbp\t\tposition\n";

#put into array;
@input = <INPUT>;


#assigning vars

$counter = 0;
$count_sub = 0;


$query1_name = $input[0];
$query1 = $input[1];
$query2_name = $input[2];
$query2 = $input[3];
$query3_name = $input[4];
$query3 = $input[5];
#can add as many as necessary as follows:
# $query4_name = $input[6];
# $query4 = $input[7];

#chomp and clear out invisible characters here
chomp $query1;
chomp $query2;
chomp $query3;

$query1 =~ s/[^\w-]//g;
$query2 =~ s/[^\w-]//g;
$query3 =~ s/[^\w-]//g;
```

```
#print "$refseq_name \n $query_name \n"; success.

#1 and 3 need to be vars to start with then split on '', not in loop.
#this explodes the strings. Already single strings.

@query1 = split ('', $query1);
@query2 = split ('', $query2);
@query3 = split ('', $query3);


#
# print "@refseq\n\n\n@queryseq";
$length1 = scalar @query1;

#print "$query1\n$query2\n"; that printed as it should have


while ($counter <= $length1)
{
        $query1nt = $query1[$counter];
        #counter is how many things in arrays
        $query2nt = $query2[$counter];
        $query3nt = $query3[$counter];




        if ($query1nt ne $query2nt && $query1nt ne "-" && $query2nt ne
"-")
        {
                ++$count_sub;
                print OUTPUT "SNP 1,2
\t$query1nt\t$query2nt\t$counter\n";
        }
        elsif ($query1nt ne $query3nt && $query1nt ne "-" && $query3nt
ne "-")
        {
                ++$count_sub;
                print OUTPUT "SNP 1,3
\t$query1nt\t$query3nt\t$counter\n";
        }
        else
        {
                ++$count_same;
        }

++$counter;
        #print OUTPUT "SNP\t\t\t$counter\t$refseqnt\t$queryseqnt\n";
}
$totallength = $count_sub + $count_same -1;

print "The number of subs is $count_sub. \n\n";
```

```
print "$totallength should equal the length of the alignment,
$length1.\n\n";

exit;

#example file 3x_example.fasta length: 63bp
#example file SNPs: 3
```

Example PIC Counter 3x input file:

```
>spec_1
CAACAAGTATTTAGTTCATCGGAATCGAAATAACAAGAATGGGGGTTTCTTTTCTCACATAAG
>spec_2
GAACAAGTATATAGTTCA---GAATCGAAATAACAAGAAGGGGGGTTTCTTTTCTCACATAAA
>spec_3
GAACAAGTATTTAGTTCATCGGAAGCGAAATAACAAGAATGGGGGTTTCTTTTCTCACATAAC
```

Example PIC Counter 3x output file:

```
                bp              position
SNP 1,2         C       G       0
SNP 1,2         T       A       10
SNP 1,3         T       G       24
SNP 1,2         T       G       39
SNP 1,2         G       A       62
```

PIC Counter 10x script:

```
#!/usr/bin/perl
#using perl version 5.18.2


#writing a program that will count subs in a fasta or txt alignment
file

#in this program, I will treat each input[1], input[3], and so on
#and put them into an array together to be parsed.

#this program will use a multiple alignment with 10 sequences aligned
in fasta format
#the first sequence is used as the reference

#note that this only counts SNPs, not insertions/deletions.

#open gene input file; needs to be in FASTA format
#pull in from command line
open(INPUT, "$ARGV[0]") || die "Can't find input file, try again; $!n";
#open output file
open (OUTPUT, '>output_10x.txt') || die "can't open output;$!n";
#label output file
print OUTPUT "\t\t\tbp\t\tposition\n";

#put into array;
@input = <INPUT>;

#assigning vars

$counter = 0;

$count_sub = 0;



$query1_name = $input[0];
$query1 = $input[1];
$query2_name = $input[2];
$query2 = $input[3];
$query3_name = $input[4];
$query3 = $input[5];
$query4_name = $input[6];
$query4 = $input[7];
$query5_name = $input[8];
$query5 = $input[9];
$query6_name = $input[10];
$query6 = $input[11];
$query7_name = $input[12];
$query7 = $input[13];
$query8_name = $input[14];
$query8= $input[15];
$query9_name = $input[16];
```

```
$query9 = $input[17];
$query10_name = $input[18];
$query10 = $input[19];

#chomp and clear out invisible characters here
chomp $query1;
chomp $query2;
chomp $query3;
chomp $query4;
chomp $query5;
chomp $query6;
chomp $query7;
chomp $query8;
chomp $query9;
chomp $query10;

$query1 =~ s/[^\w-]//g;
$query2 =~ s/[^\w-]//g;
$query3 =~ s/[^\w-]//g;
$query4 =~ s/[^\w-]//g;
$query5 =~ s/[^\w-]//g;
$query6 =~ s/[^\w-]//g;
$query7 =~ s/[^\w-]//g;
$query8 =~ s/[^\w-]//g;
$query9 =~ s/[^\w-]//g;
$query10 =~ s/[^\w-]//g;

#print "$refseq_name \n $query_name \n"; success.

#1 and 3 need to be vars to start with then split on '', not in loop.
#this explodes the strings. Already single strings.

@query1 = split ('', $query1);
@query2 = split ('', $query2);
@query3 = split ('', $query3);
@query4 = split ('', $query4);
@query5 = split ('', $query5);
@query6 = split ('', $query6);
@query7 = split ('', $query7);
@query8 = split ('', $query8);
@query9 = split ('', $query9);
@query10 = split ('', $query10);
#
# print "@refseq\n\n\n@queryseq";
$length1 = scalar @query1;


#print "$query1\n$query2\n"; that printed as it should have


while ($counter <= $length1)
{
        $query1nt = $query1[$counter];
        #counter is how many things in arrays
        $query2nt = $query2[$counter];
```

```
        $query3nt = $query3[$counter];

        $query4nt = $query4[$counter];
        $query5nt = $query5[$counter];
        $query6nt = $query6[$counter];
        $query7nt = $query7[$counter];

        $query8nt = $query8[$counter];
        $query9nt = $query9[$counter];
        $query10nt = $query10[$counter];



        if ($query1nt ne $query2nt && $query1nt ne "-" && $query2nt ne
"-")
        {
                ++$count_sub;
                print OUTPUT "SNP 1,2
\t$query1nt\t$query2nt\t$counter\n";
        }
        elsif ($query1nt ne $query3nt && $query1nt ne "-" && $query3nt
ne "-")
        {
                ++$count_sub;
                                print OUTPUT "SNP 1,3
\t$query1nt\t$query3nt\t$counter\n";

        }
        elsif ($query1nt ne $query4nt && $query1nt ne "-" && $query4nt
ne "-")
        {
                ++$count_sub;
                                print OUTPUT "SNP 1,4
\t$query1nt\t$query4nt\t$counter\n";

        }
        elsif ($query1nt ne $query5nt && $query1nt ne "-" && $query5nt
ne "-")
        {
                ++$count_sub;
                                print OUTPUT "SNP 1,5
\t$query1nt\t$query5nt\t$counter\n";

        }
        elsif ($query1nt ne $query6nt && $query1nt ne "-" && $query6nt
ne "-")
        {
                ++$count_sub;
                                print OUTPUT "SNP 1,6
\t$query1nt\t$query6nt\t$counter\n";

        }
        elsif ($query1nt ne $query7nt && $query1nt ne "-" && $query7nt
ne "-")
        {
```

```
                ++$count_sub;
                                print OUTPUT "SNP 1,7
\t$query1nt\t$query7nt\t$counter\n";

        }
        elsif ($query1nt ne $query8nt && $query1nt ne "-" && $query8nt
ne "-")
        {
                ++$count_sub;
                                print OUTPUT "SNP 1,8
\t$query1nt\t$query8nt\t$counter\n";

        }
        elsif ($query1nt ne $query9nt && $query1nt ne "-" && $query9nt
ne "-")
        {
                ++$count_sub;
                                print OUTPUT "SNP 1,9
\t$query1nt\t$query9nt\t$counter\n";

        }
        elsif ($query1nt ne $query10nt && $query1nt ne "-" && $query10nt
ne "-")
        {
                ++$count_sub;
                                print OUTPUT "SNP 1,10
\t$query1nt\t$query10nt\t$counter\n";

        }
        else
        {
                ++$count_same;
        }

++$counter;

}
$totallength = $count_sub + $count_same - 1;

print "The number of SNPs in alignment is $count_sub.\n\n";

print "The alignment length is $length1 base pairs.\n\n";

exit;

#test file 10x_example.fasta length: 63bp
#test file SNPs: 6
```

Example PIC Counter 10x input file:

```
>spec_1
CAACAAGTATTTAGTTCATCGGAATCGAAATAACAAGAATGGGGGTTTCTTTTCTCACATAAG
>spec_2
GAACAAGTATATAGTTCATCGGAATCGAAATAACAAGAATGGGGGTTTCTTTTCTCACATAAG
>spec_3
GAACAAGTATTTAGTTCATCGGAAGCGAAATAACAAGAATGGGGGTTTCTTTTCTCACATAAG
>spec_4
GAACAAGTATTTAGTTCATCGGAATCGAAATAACAAGAATGGGGGTTTCTTTTCTCACATAAG
>spec_5
GAACAAGTATTTAGTTCATCGGAATCGAAATAAAAAGAATGGGGGTTTCTTTTCTCACATAAG
>spec_6
GAACAAGTATTTAGTTCATCGGAATCGAAATAACAAGAATGGGGGTTTCTTTTCTCACATAAG
>spec_7
GAACAAGTATTTAGTTCATCGGAATCGAAATAACAAGAATGGGGGTTTCTTTTCTCACATAAG
>spec_8
GAACAAGTATTTAGTTCATCGGAATCGAAATAACAAGAATGGGGGTTTCTTTTCTCACATAAG
>spec_9
GAACAAGTATTTAGTTCATCGGAATC-----AACAAGAATGGGGTTTTCTTTTCTCACATAAG
>spec_10
GAACAAGTATTTAGTTCATCGGAATCGAAATAACAAGAATGGGGGTTTCTTTTCTCACATAAC
```

Example PIC Counter 10x output file:

```
                bp              position
SNP 1,2         C       G       0
SNP 1,2         T       A       10
SNP 1,3         T       G       24
SNP 1,5         C       A       33
SNP 1,9         G       T       44
SNP 1,10        G       C       62
```