Molecular Evolutionary and Functional Genomic Studies of *Marshallia* (Asteraceae) Utilizing Next Generation Sequencing Technology

by

Anthony Eli Melton

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
May 9, 2015

Keywords: *Marshallia*, Cytokinin Response Factor, RNA Editing, Plastid Genome, Transcriptome

Approved by

Leslie R. Goertzen, Chair, Associate Professor of Biology
Aaron M. Rashotte, Associate Professor of Biology
Scott R. Santos, Associate Professor of Biology

Abstract

Several aspects of *Marshallia* (Asteraceae) genomics have been investigated using NGS technologies. I have assembled and characterized *Marshallia* Clade III Cytokinin Response Factor genes from DNA and RNA datasets. To inform the assembly, over 50 Clade III CRF protein sequences from Asterid taxa were mined from public databases, aligned, and analyzed via MEME analyses. Two well-supported novel C-terminus motifs were identified. Expression experiments were conducted using *Marshallia caespitosa* to determine if the *Marshallia* Clade III CRF is upregulated by similar factors as *Arabidopsis* and *Solanum* Clade III CRF genes. Expression levels of the *Marshallia* Clade III CRF was only detectable after oxidative stress and cytokinin treatments, suggesting that it is expressed at very low basal levels. This gene was also found to be up-regulated by oxidative stress and by cytokinin treatment. I assembled, aligned, and compared the plastid genomes and transcriptomes for three species of *Marshallia* (*M. mohrii*, *M. obovata,* and *M. trinervia*) to identify plastid RNA editing sites. Thirty eight editing sites were identified, with 31 occurring in coding regions. Twenty four of the identified edit sites were found to occur in other taxa. Individuals of *M. mohrii* (a putative allopolyploid) exhibited decreased editing efficiency compared to diploid members of the genus included in this study. This work has extended knowledge of Clade III CRFs beyond model systems and characterized Plastid RNA editing in a non-model plant.

Acknowledgments

I owe many thanks to my advisor, Leslie R. Goertzen. He has provided me with excellent mentorship and guidance through my MS work. He has maintained an open door policy with myself and other members of the lab and has always been willing to provide assistance when necessary. I would also like to thank my committee members Aaron M. Rashotte and Scott R. Santos for their help and advice along the way. I would like to thank Paul Zwack for his assistance and guidance during my CRF work. I would like to thank my lab mates Nathan D. Hall and Curtis J. Hansen for their help. Nathan's expertise at the command line and scripting helped make my work much more manageable. Curtis' expertise in the floristics of central Alabama was very helpful during my time as the Systematic Botany TA. I would like to thank all of the professors whose classes I have taken over the last two years. I have learned so much here and will carry all that knowledge with me throughout my career. I would like to thank the wonderful friends in Auburn who have made my time here so enjoyable. I would like to thank Kat Mincey for listening to me practice my presentations and for helping me with everything along the way. And last, but not least, I would like to thank my parents for their support in my academic endeavors. Without these people, this would not have been possible.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AGE	Agarose Gel Electrophoresis
AP2	Apetala 2
BA	Benzyladenine
CAP	Contig Assembly Program
CRF	Cytokinin Response Factor
DNA	Deoxyribonucleic Acid
EF1A	Elongation Factor 1-alpha
ERF	Ethylene Resonse Factor
F.C.	Fold Change
GSI	Genetic Similarity Index
MAPk	Mitogen Activated Protein Kinase
NGS	Next-Generation Sequencing
PCR	Polymerase Chain Reaction
RNA	Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
UTR	Un-Translated Region

Chapter One. Bioinformatics Tools and the Genus *Marshallia* (Asteraceae)

Bioinformatics and NGS Technology

In the previous generation of genetics, allozymes and capillary sequencing were the primary methods of acquiring data. Allozymes are different forms of a protein caused by amino acid altering SNPs in the coding regions of genes. Use of electrophoresis to identify individuals or species with gene variants was popular during the 1970's and 1980's, but could only reveal a limited amount of variation (Gottlieb 1977, Gottlieb 1981, Soltis *et al* 1980, Soltis *et al* 1983). Advancements in DNA sequencing eventually led to the ability to sequence genes and offered much more fine-scale resolution of relationships between taxa. Sanger sequencing was invented in 1977 by Frederick Sanger *et al* and revolutionized genetics. With this technique, researchers were able to reveal the differences in nucleotide sequences that caused variation in proteins. The amount of data for phylogenetic analyses increased immensely, as each nucleotide offered a data point for analyses. Polymerase Chain Reaction (PCR) was invented in 1983 by Kary Mullis and allowed for the amplification of specific genes. This, combined with Sanger sequencing, opened up new possibilities for genetics and with the new data came new ways to analyze it. As the efficiency of Sanger sequencing and computer power increased, software were created that could assemble sequence data into full gene contigs. CAP (Contig Assembly Program; Huang and Madan 1999) was a popular assembler that utilized an alignment and quality score assessment method to assemble genes from Sanger data. These tools were effective at quantifying genetic diversity distantly related taxa, but many closely related organisms were too genetically similar for these data to generate highly resolve phylogenies.

Within the last decade, sequencing technologies has made great advancements. In 2004, 454 Life Sciences (Branford, CT, USA) (later purchased by Roche Diagnostics, Basel, Switzerland) released the 454 sequencer, which was the first high-throughput, "next" or "second" generation sequencer. This machine was able to sequence large amounts of DNA at a time and orders of magnitude more data. The 454 machine utilized beads with adapters that bound to fragmented DNA. The beads were then isolated and had solutions containing specific fluorescent-tagged nucleotides flowed over them, one at a time. A camera then detected when a nucleotide had been incorporated and then the next solution was flowed through. ILLUMINA (San Diego, CA, USA) introduced their NGS machinery in the late 2000's and have expanded their lineup to include the HiSEQ, the MiSEQ, and the NextSEQ. ILLUMINA sequencers utilize a sequence tag that "lock" fragmented DNA into place in a flow cell. The DNA fragments bind to the flow cell in a U-shaped fashion and are amplified through a PCR procedure. The new amplicons then bind to flow cell as well creating an area of high density for the fragment. Fluorescently-tagged nucleotides are then flowed through the cell as new amplicons are synthesized from both ends. As the nucleotides are incorporated they give off a fluorescent light which is picked up by a camera. Each different nucleotide gives off a different color light, which allows the machine to determine which nucleotide had been incorporated. This technique gives “paired-end” reads as each fragment will give two reads: one sequenced from the 5’ end and one from the 3’ with some small, known distance in between. While 454 Life Sciences and ILLUMINA machines can only sequence short (~100 bp) fragments of DNA, PacBio (Menlo Park, CA, USA) machines can sequence contigs of up to 5 kb in length. The PacBio sequencers use a taq-polymerase that is "locked" into position while DNA synthesis occurs in a pool of nucleotides. Sequencing occurs as light is detected by the spectrophotometer as there is a light

light emission from the included molecules. Although this type of sequencer can give you long contigs that are easier to assemble, they are often error prone (Quail *et al* 2012).

With the ability to sequence genomes, the need for greater analytical tools increased. A number of genome assemblers have been designed in recent years, including ABySS (Simpson *et al* 2009), SoapDeNovo (Li *et al* 2010, Luo *et al* 2012), Velvet (Zerbino 2010), Ray (Boisvert *et al* 2012), Trinity (Grabherr *et al* 2011), ALLPATHS-LG (Gnerre *et al* 2011, Ribeiro *et al* 2012) and PRiCE (Ruby *et al* 2013). These assemblers are designed to assemble large contigs from short (<100 bp to several kbp) DNA or RNA reads. They use a variety of different methods, including de Bruijn graphs and "greedy" algorithms that utilize alignment and mapping to extend contigs.

The assemblers Ray, Trinity, SOAP, VELVET, ABySS, and ALLPATHS all utilize a de Bruijn graph method of assembly. These programs break reads into smaller segments, known as k-mers, of a specified length of k and locates areas of overlap between them. The areas of overlap are then broken up into smaller "left" and "right" k-1 mers and the graph assembly begins. The k-1 mers become "nodes" and an "edge" is a line that connects any of the smaller left and right k-mers. Once a graph is constructed, contigs are assembled from the paths of the graph. Some methods require what is called an Eulerian walk, which refers to a path that visits every edge exactly once. This path is often considered the "solution" to assembling the k-mers, as it includes all edges connecting the k-mers in some specific order that does not repeat. This method of assembly excels at handling large amounts of data, but has difficulty assembling repeat regions. PRICE is another popular assembler, which does not use the de Bruijn graph method. Instead, it uses a "greedy" algorithm that utilizes mapping to identify reads that match a "seed", or starting sequence, and uses them to extend outwards.

The assembler Ray (Boisvert *et al* 2010) is a de Bruijn graph method assembler that is designed to assemble high-throughput read data from various sequencing technologies. While the Ray algorithm is based off previous de Bruijn graph methods, it differs by not relying on an Eulerian walk. The algorithm will stop assembly if it is not clear how to extend the current contig from the read pool. While this will decrease the overall contig lengths in the assembly, it will increase the overall accuracy.

The transcriptome assembler Trinity (Grabherr *et al* 2011) is another example of a de Bruijn graph method assembler. This assembler was designed to assemble transcripts and their variants from high-throughput cDNA sequence data. It consists of three components: *Inchworm, Chrysalis,* and *Butterfly*. *Inchworm* uses a greedy k-mer based algorithm to assemble large contigs. This process consists of six steps: 1) assembly of a k-mer dictionary from all reads within the dataset, 2) removal of low-quality k-mers from the dictionary, 3) selection of the most abundant k-mer to serve as the seed for assembly, 4) extension of the seed in both directions with the most highly occurring k-mer with k-1 overlap with the current contig terminus and concatenating it to the contig, 5) further extension of the contig in either direction until it can be extended no more and reporting the linear sequence, and 6) repeating of steps 3-5 starting with the next most highly abundant k-mer until the dictionary has been exhausted. *Chrysalis* uses a de Bruijn graph method to assemble larger components from the contigs assembled by *Inchworm* in a three step process: 1) contigs that are overlapped by k-1 are grouped into connected components, 2) a de Bruijn graph is made using the recently assembled components, 3) each read is assigned to the component with which it shares the greatest number of k-mers. *Butterfly* assembles the components into transcripts in a two stage fashion: 1) consecutive nodes within the

de Bruijn graph are merged into linear paths into nodes that represent longer sequences, 2) then, sequence termini that are weakly supported are trimmed.

Many efforts have been made to assess the quality of these different software and evaluate methods of assembly. The Assemblathon is a recurring meeting of Bioinformatists that use the latest genome assemblers and pipelines to assemble already sequenced genomes de novo and compare to a reference. The first Assemblathon (Earl *et al* 2011) was a competition to test various methods of genome assembly on a simulated ILLUMINA data set. A total of 41 genomes from 17 teams were submitted. Some of the assembly software used included SOAP de novo, ABySS, VELVET, PRICE, and ALLPATHS-LG. Assemblathon II (Bradnam *et al* 2013) assessed a number of assemblies of three vertebrate genomes. This assemblathon included many of the previous software, but also included the recently introduced software Ray. With each assemblathon, it was found that there was a great amount of variability between the different assemblies and that the best method of assembly would depend on what assessment of assembly you valued (*e.g.*, mean contig length and accuracy).

Marshallia (Shreb. 1791)

The genus *Marshallia* (Shreb. 1791) is a small southeastern U.S. endemic genus (8 species) in the sunflower family, Asteraceae (Figures 1.1 and 1.2). The flowers of *Marshallia* are typically white to pinkish-purple and occur in globose discoid heads. These pinkish and somewhat spherical inflorescences have led to a common name of "puffballs", though they are more often referred to as Barbara's Buttons. All members of the genus feature disc flowers with long, twisting corolla lobes, paleaceous receptacles, anthocyanic anthers, and simple, alternate leaves. Traditionally, these plants have had little economic value though some species, such as

M. grandiflora, have begun to make an appearance in the native plant industry due to their aesthetically pleasing nature.

Marshallia has been a bit of a taxonomic nomad, being moved from taxon to taxon, even within the higher ranks of the family. This genus has been described as "strange" (Baldwin 2009) and has been difficult to place within the family due to its unusual combination of characteristics. It has been placed in at least eight tribes within the family Asteraceae. The genus shares the characteristics of discoid head inflorescences and white to purple flowers with the Eupatorieae, a long tubular corolla with the Mutisieae, paleaceous receptacles with the Vernonieae, prenylflavonoids with the Inuleae and Heliantheae, thick green palea with the Heliantheae and Gaillardinae, and anthocyanic anthers with the Bahieae and Madieae. *Marshallia* was originally placed within the Heliantheae-Helenieae tribal complex (Lessing 1832, De Candolle 1836, Torrey and Gray 1942, Bentham and Hooker 1873) but was later excluded by Stuessy (1977). Turner and Powell (1977) found that the genus should fall within the Eupatoreae, based on similarity with the genus *Palafoxia*. As molecular techniques improved, our understanding of the higher relationships within the Asteraceae improved. In 1981, Robinson rejected the reclassification of Turner and Powell (1977) and maintained *Marshallia* into a monogeneric subtribe, the Marshallinae, within the Heliantheae tribe, though he hypothesized that the genus may fall within the tribe Inuleae *s.l.* Since the mid-1990s, it has been shown that the Heleneae is a clade that falls within the Heliantheae *sensu latu* and that *Marshallia* belongs to its own clade within the Heleneae (Kim and Jansen 1995, Baldwin *et al* 2002, Goertzen *et al* 2003, Reveal 2012).

In addition to morphological and molecular techniques, chemical analyses have been conducted to determine the placement of *Marshallia*. In 1980, Bohlmann *et al* determined that

the genus shared a number of chemical derivatives with members of the Inuleae. Herz and Bruno (1987) characterized a dozen novel chemical derivatives from *M. graminifolia* and found that they were not shared with members of the Heliantheae. In 1988, Jakupovic *et al* expanded on the previous work and added chemical data from six *Marshallia* taxa. Based on the results of the chemical analyses, the authors concluded that the genus did not belong in the Heliantheae, but should be placed somewhere intermediate of the Heliantheae and Inuleae.

Marshallia species exhibit high levels of similarity and can be difficult to differentiate. Characters that can vary between species include single or multiple capituleae and habitat. The members of this genus vary greatly in their distribution and abundance. Three species are common, one is uncommon, two are rare, one is federally listed as threatened, and one was recently described with only two known populations.

M. caespitosa Nutt. ex DC (Figure 1.3) is a common species that occurs in limestone outcrops and prairies in eastern Texas, western Louisiana, and northwards to Missouri and Kansas. It typically has white flowers like *M. obovata*, but can have branched peduncles in var. *signata*. This species has two described subspecies: *M. caespitosa* var. *caespitosa* Nutt. ex DC and *M. caespitosa* Nutt. ex DC var. *signata* Beadle & F.E. Boynt. *M. caespitosa* var. *caespitosa* has been found to be a tetraploid, while var. *signata* is a diploid, but little is known about its genome evolution.

M. graminifolia (Walt.) Small (Figure 1.4) is a common bog species that ranges from east Texas eastwards into North Carolina. It has two described varieties that have been variously treated over the years, *M.graminifolia* var. *graminifolia* (Walt.) Small and *M. graminifolia* var. *cyanthera* (Ell.) Beadle & F.E. Boynt.

M. grandiflora Beadle & F.E. Boynt (Figure 1.5) is state-listed as rare in Tennessee, West Virginia, and Pennsylvania and occurs along riverbanks and shoals in Tennessee and West Virginia. This species epithet comes from it having the largest flowers of the *Marshallia.*

M. legrandii Weakley and Pointdexter (Figure 1.6) has only two known extant populations, one in North Carolina and one in Virginia and occurs in mafic woodlands and prairies. It is named after the botanist Harry E. LeGrand, who, in 1986, discovered a strange population of what he thought to be *M. grandiflora*. This population's members had an unbranched peduncle, basal leaves, pink to deep pink corollas, and pubescence on the stem below the inflorescence. It was later described as a new species by Weakley and Pointdexter in 2012. It is also the tallest of the *Marshallia*, reaching up to 9dm in height.

M. mohrii Beadle & F.E. Boynt (Figure 1.7) is a federally listed threatened species that occurs on granite outcrops in central to northern Alabama and part of west Georgia. It is a polyploid species that is proposed to have evolved from hybridization between *M. trinervia* and *M. grandiflora* (Watson *et al* 1991).

M. obovata (Walt.) Beadle & F.E. Boynt (Figure 1.8) is a common species that occurs in sandy pine woods from eastern Alabama northwards to West Virginia. Its epithet stems from its obovate paleae. Its flowers are white and individuals of the species typically have an unbranched peduncles. It has two described variations, *M. obovata* var. *obovata* (Walt.) Beadle & F.E. Boynt and *M. obovata* (Walt.) Beadle & F.E. Boynt var. *scaposa* Channell which can be differentiated by whether the leaves are basal (*M. obovata* var. *scaposa*) or cauline (*M. obovata* var. *obovata*).

M. ramosa Beadle & F.E. Boynt (Figure 1.9) is state-listed as rare in Georgia and endangered in Florida and occurs in sandstone outcrops and in isolated populations Georgia, with

few populations occurring in northern Florida and on the AL-GA border. It can have highly branched peduncles which lead to the specific epithet "ramosa", meaning branched.

M. trinervia (Walt.) Trel (Figure 1.10) is an uncommon species that occurs along riverbanks and shoals and has populations spread throughout Alabama, Mississippi, and Tennessee, with few populations occurring in Louisiana and Georgia, and one putative population in eastern South Carolina. It was named "trinervia" because its leaves have three prominent nerves that run from their bases to tips.

In 1957, a phenetic analysis conducted by Chanell found that there were four complexes, including: the Grandiflora complex (*M. grandiflora*, *M. mohrii*, and *M. trinervia*), the Obovata complex (*M. obovata* varieties), the Caespitosa complex (*M. caespitosa* varieties and *M. ramosa*), and the Graminifolia complex (*M. graminifolia* and *M. tenuifolia*). A phenetic analysis in 1990 by Watson and Estes recovered four complexes as well, but were not in agreement with those of Chanell's. *M. grandiflora* and *M. trinervia* were again clustered together, but, in disagreement with Chanell's work, with *M. mohrii* being grouped with the *M. graminifolia* complex. *M. caespitosa* and *M. obovata* were clustered together while *M. trinervia* was un-clustered.

The members of this genus, such as *M. graminifolia*, have been variously described and have been resistant to traditional classification techniques. It has been found that morphological or traditional molecular techniques are not sufficient to resolve the phylogeny of this genus. Various molecular phylogenetic analyses conducted in the 90's by Watson *et al* found that this genus is not only very morphologically similar, but also genetically similar. Populations of *M. graminifolia*, which has been treated as two species and one species with two subspecies or varieties, were found to exhibit a genetic similarity index (GSI) of 0.99. This greatly exceed the

typical GSI of conspecifics. Adding in another level of complexity to the phylogeny of this group are the occurrence of two polyploids, *M. mohrii* and *M. caespitosa* var. *caespitosa*.

While no work has been conducted on the origins of polyploidy in *M. caespitosa*, some efforts have been made to elucidate the origins of *M. mohrii*. Watson *et al* found that there were shared alleles between *M. mohrii, M. trinervia,* and *M. grandiflora*, suggesting that *M. trinervia* and *M. grandiflora* may be involved in the hybridization events that led to the evolution of *M. mohrii*. While there were shared alleles found, there were also a number of alleles that were unique to *M. mohrii*. These results were in agreement with phylogenies based on morphology by Chanell in the 1950's that suggested a relationship between the three, but also suggest that there may be a third species involved (Watson *et al* 1991).

Due to the taxonomic history and phylogenetic troubles of *Marshallia*, it is a perfect example of a non-model organism that is fit to be researched using NGS technologies towards resolution of phylogeny. Large quantities of data will be needed to resolve the phylogeny of the genus and answer questions about the origins of its polyploid species. For this reason, large data sets of DNA and RNA designed for genome-skimming analyses have been compiled using ILLUMINA sequencing data. The primary goal of this project is to utilize these data sets and evaluate their useful in contexts for which NGS data would be novel. These datasets have been used in the assembly, description and characterization of single copy nuclear gene, which would typically be done using PCR and Sanger sequencing. The plastid genomes and transcriptomes of several species of *Marshallia* have been assembled and compared to identify RNA editing sites. This work has been traditionally conducted using PCR to amplify DNA and cDNA for comparison. This approach has streamlined the process and made many edit sites obvious and has allowed for the identification of novel sites within and outside of coding regions.

Literature Cited

1. Beadle, C.D. and F.E. Boynton. 1901. Revision of the species of *Marshallia*. Biltmore Bot. Studies 1: 3–10.

2. Baldwin, B.G. 2009. Heliantheae alliance. Chapter 41, *in* V.A. Funk, A. Susanna, T.F. Stuessy, and R.J. Bayer (eds.). Systematics, Evolution, and Biogeography of Compositae. International Association for Plant Taxonomy, Vienna, Austria.

3. Bentham and Hooker. 1873. Genera Plantarum 2. 198Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., & Corbeil, J. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biology*, *13*(12), R122. doi:10.1186/gb-2012-13-12-r122

4. Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., … Korf, I. F. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1), 10. doi:10.1186/2047-217X-2-10

5. Channell, Robert Bennie 1955. A revisional study of the genus *Marshallia* Schreb (Compositae). Thesis (Ph. D.) Duke University.

6. De Candolle, A. P. 1836. *Marshallia* in Podromus systematis naturalis regni vegetabilis 5: 680. Paris

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

8. Gottlieb, L. D. 1977. Electrophoretic Evidence and Plant Systematics. *Annals of the Missouri Botanical Garden*, *64*(2), 161–180. http://doi.org/10.2307/2395330

9. Gottlieb, L D. 1981. Electrophoretic evidence and plant populations. Prog Phytochem, 7, 1–47.

10. Gnerre S, MacCallum I, Przybylski D, Ribeiro F, Burton J, Walker B, Sharpe T, Hall G, Shea T, Sykes S, Berlin A, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. High-quality draft assemblies of mammalian genomes from massively parallel sequence data *Proceedings of the National Academy of Sciences USA* (January 2011 vol. 108 no. 4 1513-1518).

11. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. doi:10.1038/nbt.1883

12. Haridas, S., Breuill, C., Bohlmann, J., & Hsiang, T. 2011. A biologist's guide to de novo genome assembly using next-generation sequence data: A test with fungal genomes. *Journal of Microbiological Methods*, *86*(3), 368–375. doi:10.1016/j.mimet.2011.06.019

13. Hoffman . 1890. Compositae in Engler & Prantl *Die naturlichen Pflantzenfamilien* IV. 5: 24-247.

14. Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. Genome Res., 9, 868-877.

15. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., … Wang, J. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*(2), 265–272. doi:10.1101/gr.097261.109

16. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18. doi:10.1186/2047-217X-1-18

22. Reveal, J. L. 2012. An outline of a classification scheme for extant flowering plants. Phytoneuron 2012-37: 1–221. Published 23 April 2012. ISSN 2153 733X

23. Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., … Gu, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*(1), 341. http://doi.org/10.1186/1471-2164-13-341

24. Ribeiro F, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, Berlin AM, Montmayeur A, Shea TP, Walker BJ, Young SK, Russ C, MacCallum I, Nusbaum C, Jaffe DB. 2012. Finished bacterial genomes from shotgun sequence data. *Genome Research* **22:** 2270–7.

25. Ruby, J. G., Bellare, P., & Derisi, J. L. 2013. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda, Md.)*, *3*(5), 865–880. http://doi.org/10.1534/g3.113.005967

26. Sanger, F., Nicklen, S., & Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463-5467.

27. Schatz, M. C., Delcher, A. L., & Salzberg, S. L. 2010. Assembly of large genomes using second-generation sequencing. *Genome Research*, *20*(9), 1165–1173. doi:10.1101/gr.101360.109

28. Schreber, Johann Christian Daniel von. 1791. Genera Plantarum 2: 810

29. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, İ. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123. doi:10.1101/gr.089532.108

30. Small, J. K. 1901. Marshallia in Flora of the southeastern United States 1283-1285. New York

31. Soltis, D. E., Haufler, C. H., Darrow, D. C., & Gastony, G. J. 1983. Starch Gel Electrophoresis of Ferns: A Compilation of Grinding Buffers, Gel and Electrode Buffers, and Staining Schedules. *American Fern Journal*, *73*(1), 9–27. http://doi.org/10.2307/1546611

32. Soltis, D. E., Haufler, C. H., & Gastony, G. J. 1980. Detecting Enzyme Variation in the Fern Genus Bommeria: An Analysis of Methodology. *Systematic Botany*, *5*(1), 30–38. http://doi.org/10.2307/2418733

33. Torrey, J. and A. Gray. 1842. *Marshallia* in Flora of North America 2:390-391. New York.

34. Voelkerding, K. V., Dames, S. A., & Durtschi, J. D. 2009. Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry*, *55*(4), 641–658. doi:10.1373/clinchem.2008.112789

35. Watson, L. E., Elisens, W. J., & Estes, J. R. 1991. Electrophoretic and Cytogenetic Evidence for Allopolyploid Origin of Marshallia mohrii (Asteraceae). *American Journal of Botany*, *78*(3), 408–416. doi:10.2307/2444963

36. Watson, L. E., Elisens, W. J., & Estes, J. R. 1994. Genetic variation within and among populations of the Marshallia graminifolia complex (Asteraceae). *Biochemical Systematics and Ecology*, *22*(6), 577–582. doi:10.1016/0305-1978(94)90069-8

37. Watson, L. E., & Estes, J. R. 1990. Biosystematic and Phenetic Analysis of Marshallia (Asteraceae). *Systematic Botany*, *15*(3), 403–414. doi:10.2307/2419354

38. Watson, L. E., Jansen, R. K., & Estes, J. R. 1991. Tribal Placement of Marshallia (Asteraceae) Using Chloroplast DNA Restriction Site Mapping. *American Journal of Botany*, *78*(8), 1028–1035. doi:10.2307/2444891

39. Weakley, A.S. and D.B. Poindexter. 2012. A new species of Marshallia (Asteraceae, Helenieae, Marshalliinae) from mafic woodlands and barrens of North Carolina and Virginia. Phytoneuron 2012-105: 1–17. Published 26 November 2012. ISSN 2153 733X

40. Zerbino, D. R. 2010. Using the Velvet *de novo* assembler for short-read sequencing technologies. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *CHAPTER*, Unit–11.5. doi:10.1002/0471250953.bi1105s31

Tables and Figures

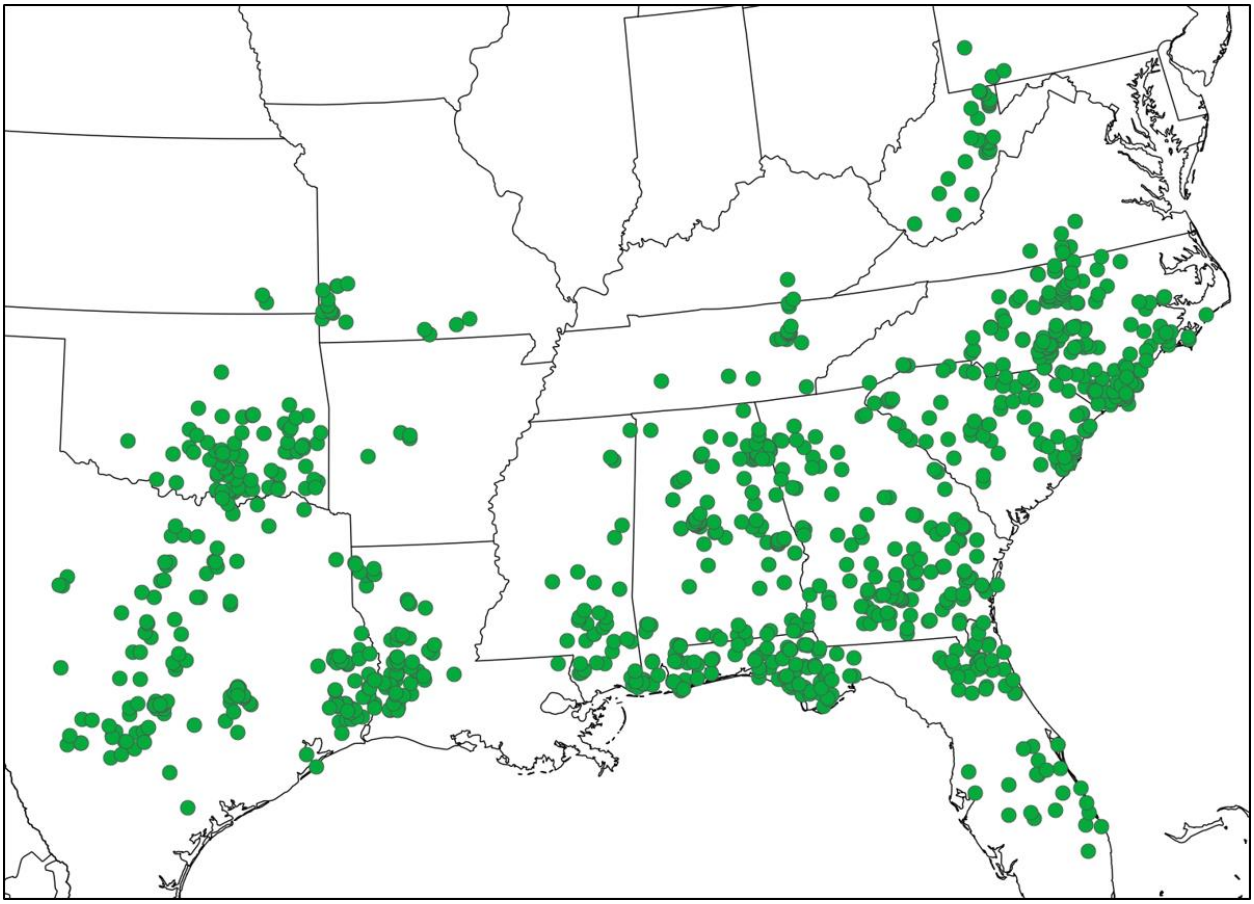

Figure 1.1. Overall distribution of *Marshallia*. Compliments of Curtis J. Hansen.

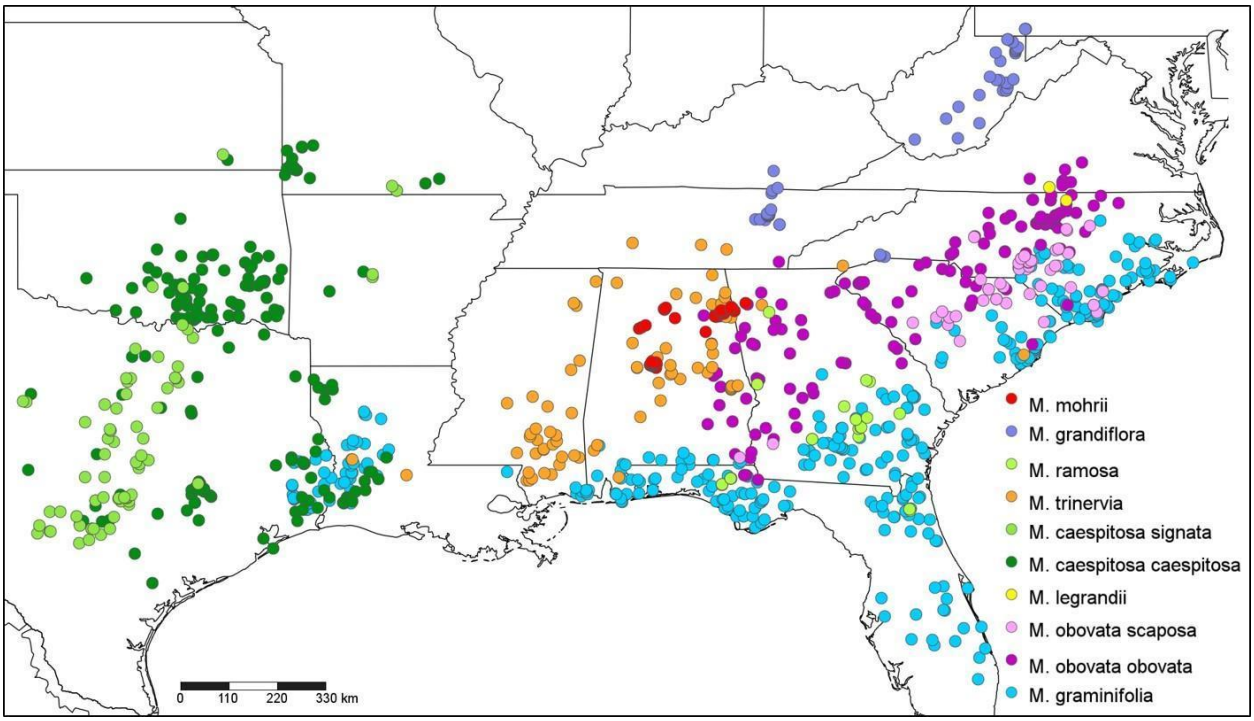

Figure 1.2. Distributions for *Marshallia* species. Compliments of Curtis J. Hansen.

Figure 1.3. *M. caespitosa*, photo compliments of Curtis J. Hansen.

Figure 1.4. *M. graminifolia*, photo compliments of Curtis J. Hansen.

Figure 1.5. *M. grandiflora*, photo compliments of Curtis J. Hansen.

Figure 1.6. *M. legrandii*, photo compliments of Curtis J. Hansen.

Figure 1.7. *M. mohrii*, photo compliments of Curtis J. Hansen.

Figure 1.8. *M. obovata*, photo compliments of Curtis J. Hansen.

Figure 1.9. *M. ramosa*, photo compliments of Curtis J. Hansen.

Figure 1.10. *M. trinervia*, photo compliments of Curtis J. Hansen.

Chapter Two. Sequencing and Functional Characterization of the *Marshallia* (Asteraceae) Clade III Cytokinin Response Factor (CRF)

Abstract

Cytokinin Response Factor (CRF) genes are a subgroup of AP2/ERF domain containing transcription factors that are up-regulated by cytokinin. These genes are defined by the CRF and AP2/ERF domains and have been phylogenetically separated into five clades. Clade III CRFs have been found to respond to oxidative stress and other abiotic factors as well as cytokinin induction in *Arabidopsis* and *Solanum lycopersicum*. In this study, over 50 Asterid Clade III CRF protein sequences were aligned and analyzed for unique, clade III defining motifs. MEME analysis identified a novel Clade III CRF unique C-terminus motif. The Clade III CRF genes of *Marshallia mohrii* and *M. caespitosa* were assembled from whole genomic and transcriptomic data collected and analyzed via NGS technology and assembly methods. The *M. caespitosa* Clade III CRF gene demonstrated strong sequence conservation in the functional domains when compared to *SlCRF5*, another Asterid Clade III CRF, and the only *S. lycopersicum* Clade III CRF. Asterid 5' UTRs were mined from public databases and aligned to identify conserved promoter motifs that might function in Clade III CRF regulation. To better understand functional conservation of the Clade III CRFs, expression experiments were conducted on *M. caespitosa*. Very low levels of basal expression were detected for the *M. caespitosa* Clade III CRF, with up-regulation occurring in both oxidative stress and cytokinin treatment. These results suggest there is conservation in the sequence, form, and regulatory mechanisms of Clade III CRF genes and proteins.

Introduction

Cytokinin is a plant hormone that functions in cell proliferation and differentiation (Mok and Mok 2001). It was studied by Folke Skoog in the 1940's and was described in more detail by Miller *et al* in the 1950's (Miller *et al* 1955, Miller *et al* 1956). It was first called kinetin and was thought to be derived from DNA. Some early research involved tobacco callus cultures and it was found that cultures treated with kinetin showed an increase in mass over control groups, suggesting a stimulation of cell division. Evidence to determine the derivation of kinetin came when Miller *et al* (1956) conducted numerous experiments to reveal its structure. Hydrolysis reactions and analysis of the products via UV-vis Spectroscopy produced results that resembled that of adenine, a nitrogenous base of DNA. Kinetin also shared the behavior of adenine when analyzed with column chromatography. These results, combined with previous experiments, allowed the researchers to hypothesize the structure of kinetin. Since then, this hormone has been found to function in cell differentiation, vascular differentiation, seed development, and plastid division. It has also been found to play a role in several other aspects of plant biology, including nutrient balance, stress response and leaf senescence (Okazaki *et al* 2009, Muller and Sheen 2007) as well as gene regulation (Schmulling*et al* 1997).

Natural cytokinins are adenine derivatives and are classified by the configuration of their N6-side chain as isoprenoid or aromatic cytokinins. There are a variety of natural cytokinins, including: zeatins, dihydrozeatin, N^6 benzyladenine (BA), and topolin. Zeatins are derived from adenine by hydroxylation. Dihydrozeatin is an adenine derivative that features a saturated side chain. Topolin is an adenine derivative that features an aromatic side-chain (Strnad *et* al 1997). They can be converted between cytokinin bases, nucleosides and nucleotides (Mok and Mok 2001). Cytokinin has been found to function with auxin, stimulating cell division (Miller *et al*

1955). The cytokinin to auxin ratio influences cell division, with higher ratios favoring development of the shoot while lower ratios favor the development of the roots (Skoog and Miller 1957).

Cytokinin induction in the model plant *Arabidopsis* occurs through binding of cytokinin to sensor histidine kinases (AHKs), such as CKI1 and CRE1/AHK4/WOL, which are hybrid histidine kinases that feature a conserved receiver domain. This ligand binding initiates a modified two-component phosphorelay in which *Arabidopsis* histidine-containing phosphotransfer proteins (AHPs) are phosphorylated. The phosphorylation of an AHP occurs at a conserved histidine residue and is typically restricted to the cytoplasm. After phosphorylation in the cytoplasm, these proteins are translocated to the nucleus where they transfer the phosphate to the *Arabidopsis* response regulators (ARRs), which are a type of transcription regulator. There are two types of ARRs (cytokinin-inducible A-ARRs, and DNA-bind B-ARRs), which affect transcription in opposite ways. Type-A ARR's are transcription repressors, while Type-B ARRs are transcription activators. (Hwang and Sheen 2001).

Cytokinin Response Factors are a subfamily of AP2/ERF domain containing genes that are up regulated by cytokinin (Rashotte *et al* 2006, Zwack *et al* 2012). These genes were first discovered in experiments of cytokinin response in *Arabidopsis.* It was found that some these genes were highly induced by cytokinin induction (Rashotte *et al* 2003). These transcription factor coding genes have been found to contain several conserved domains, including: the CRF domain, the AP2/ERF domain, and a putative kinase domain, MAPk (Rashotte and Goertzen, 2010). CRF proteins have been found to form homo- and hetero- dimers with other CRFs, most likely at the CRF domain. They were also found to interact primarily with AHPs, and rarely with AHKs, nor type-A or type-B RR proteins (Cutcliffe *et al* 2011). This suggests that the CRF-CRF

and CRF-AHP interactions may be significant in cytokinin-regulated developmental processes. The AP2/ERF domain is a DNA binding domain that allows the protein to bind to various DNA sequences and regulate transcription (Sakuma *et al* 2003). This domain was considered to present only in land plants, but weak homologs have been identified in bacteria and viruses (Magnani *et al* 2004). The function of the putative MAPk domain in CRFs has not been experimentally determined but the sequence is very similar to that of known MAPk domains found in a broad range of land plants (Rodriguez *et al* 2010).

Phylogenetic analyses using the CRF and AP2/ERF domains of CRF proteins across a wide range of taxa have indicated that there are at least five major clades of CRFs in flowering plants. Each clade features a unique C-terminus motif in the protein sequence (Zwack *et al* 2012). Genes within individual clades have been found to be involved in a range of different functions and are regulated by various mechanisms. Clades I-IV have been found to be most highly expressed in the phloem tissue of aerial organs, suggesting they may function in cytokinin-regulated development processes within these organs and tissues (Zwack *et al* 2012). Clade II CRFs, *AtCRF3* and *At*CRF4, have been found to be unaffected by cytokinin, although they still contain the conserved domains of the CRF proteins (Rashotte *et al* 2006). Some of the *Arabidopsis* CRFs, particularly those of Clade III, have been found to influence leaf vasculature patterning, plant growth and development, and leaf senescence (Zwack *et al* 2013).

Arabidopsis thaliana has been the primary organism used in the study of CRFs. 12 CRF genes have been identified in *Arabidopsis*. The genes *At*CRF1 and *At*CRF2 have been found to be Clade I CRF genes. There are two Clade II *Arabidopsis* CRF genes, *At*CRF3 and *At*CRF4. *Arabidopsis* has two Clade III CRF genes: *At*CRF5 and *At*CRF6. Within Clade V, *Arabidopsis* has four genes: *At*CRF9, *At*CRF10, *At*CRF11, and *At*CRF12. *Arabidopsis* has no Clade IV CRF,

despite Clade III CRFs being present throughout the Angiosperms. This loss seems to be shared among other closely related taxa, as other members of Brassicaeae do not have any genes of this clade, although other families in the order Brassicales have been found to contain a Clade IV CRF gene. This suggests a loss of this clade's genes as the Brassicaceae began to diversify. *At*CRFs 7 and 8 are not placed in any recognized clade of CRFs, due to their lack of the MAPk domain and other identifying C-terminus elements.

Solanum lycopersicum has also been a valuable experimental system in the study of the CRFs. 11 CRF genes have been identified, with several being homologous to *Arabidopsis* CRFs. *Sl*CRF1 is a Clade IV gene, *Sl*CRF2 is a Clade I gene, s*Sl*CRF4 and *Sl*CRF6 are of Clade II, *Sl*CRF5 is a Clade III gene, and *Sl*CRFs 9, 10, and 11 are of Clade V. Like *At*CRF3 and *At*CRF4, *Sl*CRF4 and *Sl*CRF6 do not respond to cytokinin but do contain CRF domains (Shi *et al* 2012). The other *Sl*CRFs do not fit into any of the recognized clades.

The CRF genes that do not fit into one of the five clades feature unique domain patterns and do not appear to have orthologues in *Arabidopsis*. *Sl*CRF3 contains two CRF domains and an AP2/ERF domain, in an alternating order. This is unique among the CRFs as all others characterized have had only one CRF domain. *Sl*CRFs 7 and 8 have been found to be truncated with respects to the typical CRF protein content. They have the CRF and AP2/ERF domains but lack the MAPk domain and all downstream elements, similar to *At*CRFs 7 and 8.

The Clade III CRFs have been found to function in cytokinin induced senescence of leaves, oxidative stress response, and response to other abiotic stresses, such as cold stress (Shi e*t a*l 2012, Zwack *et al* 2012, Zwack *et al* 2013, Gupta, 2013). *At*CRF6 (an *Arabidopsis* Clade III CRF) is thought to function downstream of the Two-Component Cytokinin Signaling Pathway in negative regulation of leaf senescence (Zwack *et al* 2013). *Sl*CRF5 (a tomato Clade III CRF) has

been found to be most highly expressed in root and leaf tissue where it is thought to function in growth and development. It has also been found to be induced by flood and drought stress, oxidative stress, ABA stress, and low temperature stress in roots (Gupta 2013).

Research focusing on gene sequencing and characterization have been most often approached using traditional molecular techniques, such as PCR and capillary sequencing. Although these techniques have been successful across a wide range of disciplines, there have been great advances in sequencing technology in recent years. Next-generation platforms have given us the ability to produce genome-scale data with relative ease and low cost. These technologies have been utilized in research ranging from sequencing and assembly of PCR amplicons to entire genomes. This work will serve to examine the feasibility of a genome-skimming approach in the context of gene description and characterization.

The goals of the research described here were to identify conserved motifs unique to Clade III CRF proteins, to sequence and characterized a Clade III CRF gene for the genus *Marshallia* (Shreb.) and to test for conservation of regulatory mechanisms identified in *Arabidopsi*s and *Solanum* Clade III CRFs in the *Marshallia* Clade III CRF gene. This work will help increase our knowledge of Clade III CRF form and regulatory mechanisms significant to the Clade III CRF proteins.

Methods

Asterid Clade III CRF data

Asterid Clade III CRF protein sequences were collected from public databases (*e.g.*, NCBI (Benson *et al* 2005), OneKP (One Thousand Plant Project), and Phytozome (Joint Genome Institute, Walnut Creek, CA)) and aligned manually in SeaView v. 4.4.2 (Gouy *et al*

2010). Sequences of low quality (*e.g*., missing large regions, short sequences, etc.) were excluded from further analysis. ClustalO v. 1.2.0 (Sievers *et al* 2011) was used to align the remaining sequences to compare to manual alignment. Fifty unaligned sequences were submitted to MEME motif analysis (Bailey and Elkin 1994). Parameters for MEME analysis were set as follows: occurrence of motif = zero to one, number of motifs = 10, minimum number sites = 35, maximum number sites = 50, minimum width = five, maximum width = 100.

Extraction and Sequencing

DNA from all taxa of *Marshallia* were extracted from fresh leaf tissue using a modified 2X CTAB protocol as described by Doyle and Doyle (1987) or E.Z.N.A. kits (Omega Bio-tek, Inc., Norcross, GA) per manufacturer protocol. RNA were extracted from fresh leaf material of five individuals of *Marshallia* (two *M. mohrii*, one *M. caespitosa*, one *M. obovata*, and one *M. trinervia*) and from fresh shoot and roots of *M. caespitosa* using Plant RNA extraction kit (Qiagen, Hilden, Germany) per manufacturer protocol. DNA were submitted to HudsonAlpha Institute for Biotechnology (Huntsville, AL) for paired-end library prep and 100bp sequencing via an ILLUMINA (ILLUMINA Inc., San Diego, CA) HiSEQ 2000 platform. RNA samples were submitted to the Auburn University Genomics and Sequencing Laboratory (Auburn, AL) where cDNA libraries were prepared using an Illumina mRNA TruSeq kit, which uses mRNA beads with poly-T sequences to isolate mRNA. Sequencing was performed using an ILLUMINA HiSEQ 1500 platform.

Assembly and Description

For the coding region of Clade III CRF gene, DNA reads from all unassembled DNA read sets were identified using BLAST (Altschul *et al* 1990) and assembled using CAP3 (Huang

and Madan 1999). RNA reads from *Marshallia* accessions M2.9 (*M. trinervia*), M3.9 (*M. obovata*), M10.1.1 (*M. caespitosa*), M20r, and M21r (*M. mohrii*) were assembled using Trinity (Grabher *et al* 2011). Both DNA and RNA reads were mapped to the resulting DNA contig and RNA transcript using Bowtie2 v. 2.1.0 with the flags --local, --qc-filter, and --no-unal (Langmead and Salzberg 2012). Maps were visualized using Tablet 1.14.04.10 (Milne *et al* 2013) to assess quality of the mapping and to assess accuracy of the assemblies.

5' UTR sequence assembly was performed by identifying appropriate reads from all data sets via BLAST followed by hand assembly in SeaView. Bowtie2 was used to map read pairs (pairing enforced mapping) from all data sets to further identify the reads of the 5' UTR. Within the mapping, reads were selected that were properly mapped, along with their pair, and that exhibited soft clipping of the 5' region. These reads were then extracted from the database and added to the assembly.

Characterization of Expression

Individuals of *Marshallia caespitosa* were grown in the Auburn University Plant Research Center greenhouses. Cypselae were planted in Sunshine (Sun Gro Horticulture, Agawam, MA) mix #8 potting soil just below the surface of the soil and covered with a thin layer peat moss. A weekly watering/fertilizing regime and natural light and photoperiod were utilized. Plants were grown for approximately one year and were harvested while in the rosette growth stage. Whole plants were used in oxidative stress and cytokinin induction tests. Three plants were sprayed with a control spray, three plants were sprayed with 0.5 μM solution of benzyladenine (BA), and three were sprayed with a 0.5% solution of hydrogen peroxide. Plants were then given a six hour period to allow for uptake of the solutions and for change in expression to occur. Each plant was then separated into shoots and shoots and RNA was

extracted from the tissues using a Qiagen Plant RNA extraction kit (Qiagen, Hilden, Germany) per manufacturer protocol. A cDNA library was prepared using Quanta qScript cDNA supermix (Quanta BioSciences, Inc., Gaitherburg, MD) per manufacturer protocol with an Eppendorf Thermal Cycler (Eppendorf, Hamburg, Germany). PCR was performed in an Eppendorf Thermal Cycler with G-BIO Taq per modified protocol (Table 2.1). PCR was conducted on various constitutively expressed genes to evaluate their use as a control in qPCR measurements. α-tubulin, π-tubulin, elongation factor 1-alpha, and actin were amplified following the previously described methods and compared by agarose gel electrophoresis. EF1A was selected as the control gene for qPCR based on consistent banding in all test groups. q-PCR was performed using two technical replicates of three biological replicates per experimental group in an Eppendorf Realplex with Quanta Sybr Taq mix per a modified protocol (Table 2.2). Primers for PCR were designed using NCBI BLAST Primer and were prepared by Eurofins Scientific (see table 2.3 for primer sequences). Mean Ct was calculated for each biological replicated. Mean Ct of EF1A were subtracted from mean Ct of the respective experimental replicate to calculate ΔCt. ΔΔCt was calculated by subtracting ΔCt of the experimental group from the ΔCt of the control group. Fold change (F.C.) was calculated by the equation 2^ΔΔCt.

Results

Sequence data for 50 Asterid Clade III CRF proteins were aligned and submitted for the MEME analysis. MEME analysis recovered several conserved motifs, including two novel Clade III specific motifs. The most highly supported of the identified motifs was present in 48 sequences, and was supported by an e-value of 2.2e-586. This C-terminus motif was 22 residues in length. The motif sequence was [LY]D[QS]CFL[NK][DE][FY]FDFRSPSP[LI][IM]Y[ED]E (Figure 2.1). Another Clade III CRF motif identified was found at the end of the C-terminus, and

consisted of a strongly conserved WDV[DN]DF[FL] sequence (e-value=9.3e-06; 36/50 sequences) (Figure 2.2). The CRF domain motif was found to be VRI[SY]VTD[CG]DATD (e-value=7.3e-066) (Figure 2.3). Three motifs were identified for the AP2/ERF domain. The most statistically supported with an e-value of 5.4e-221 and occurred third within the AP2/ERF domain (Figure 2.4). The second most statistically supported AP2/ERF domain motif had a calculated e-value of 5.2e-204 and occurred second in the AP2/ERF domain (Figure 2.5). The least supported AP2/ERF domain motif had a calculated e-value of 6.5e-200 and occurred at the beginning of the domain (Figure 2.6). The putative MAPk motif sequence was found to be SPTSVLRFD, with an e-value of 6.1-054, present in 35 sequences (Figure 2.7).

Assembly from DNA data was difficult due to low coverage. CAP3 (Huang and Madan 1999) contigs were typically truncated, containing ≤98% of the total coding region. Trinity successfully assembled a Clade III CRF transcript from start (with some upstream elements) to stop (with some downstream elements) from the M21r RNA read set (henceforth informally referred to as *Mm*CRF3; refer to Table 2.4 for coding sequence). Trinity assembled a near-complete transcript from the M10.1.1 RNA read set, from just downstream of the CRF domain to just downstream of the stop codon. The top BLAST hit from Trinity contig file provided a strong alignment and contained all elements of Clade III CRFs. The *M. mohrii* Trinity transcript was 846 nucleotides in length from start to stop codon and were translated into a 281 amino acid protein. To extend the *M. caespitosa* transcript, *M. caespitosa* DNA read data from were mapped to the M21r transcript via previously described methods and a majority-rule consensus was called (henceforth informally referred to as *Mc*CRF3; refer to Table 2.5 for coding sequence). The resulting consensus was in agreement with the Trinity transcript (where overlap occurred)

and increased the sequence to the start codon. Other accessions provided only a truncated transcript (M20r) or no assembly at all (M2.9 and M3.9).

For mapping of DNA reads to evaluate accuracy and read depth, both individual and species readsets were mapped to the *Mm*CRF3 transcript. Species readsets were created by concatenating data from each accession of each species into one dataset. Mapping of individual readsets offered low and incomplete coverage, with ~10-20 reads being mapped per individual, with an average coverage of ~1 read and average max read depth of ~3 (Table 2.6). Mapping of species readsets offered ~40-50 reads per species with an average read depth of ~2 reads and average max read depth of ~6 reads (Table 2.7). Mapping of M21r RNA read data provided much stronger coverage, with 519 reads being mapped providing 100% coverage, with an average coverage = 48.435, max coverage = 78, and only 3% mismatch.

To compare the read coverage in the Clade III CRF maps with other genes, species specific read data were aligned to actin-2, α-tubulin, π-tubulin, and elongation factor 1-alpha (EF1A) contigs. Mappings of these genes show generally higher number of reads mapped and read coverage, ranging from ~50 to ~150 reads per dataset mapping to the transcript with an average depth of ~5 (Table 2.8).

Both *Mc*CRF3 and *Mm*CRF3 exhibited some sequence variation in the conserved domains of the protein (refer to Figure 2.8 for mapping of conserved domains and motifs within protein). When compared to *Sl*CRF5, the CRF domain of *Mc*CRF3 varied in 18 positions while *Mm*CRF3 varied in 15 positions, with strong conservation of the core of the motif, D[CG]DATDDD (Figure 2.59). Within the most conserved region of the C-terminus motif, the *Mm*CRF3 protein was found to differ in three positions and the *Mc*CRF3 protein varied in four. The core of the motif was strongly conserved, sharing a DF[FL]DFR[SI]PSP[LI]M pattern

(Figure 2.10). The MAPk domain showed no variation within its domain, exhibiting the typical SP[TV]SVL motif, with a T in the 3rd position (Figure 2.11). The small motif at the end of the C-terminus varied in four out of five positions, from KWAND in *SlCRF*5 to EWLDD in both *Marshallia* Clade III CRFs (Figure 2.12). *Mc*CRF3 and *Mm*CRF3 featured 16 SNPs relative to each other, with 10 resulting in amino acids changes (Tables 2.9 and 2.10). NCBI BLASTn was used to calculate percent similarity between *Mc*CRF3, *Mm*CRF3 and *Sl*CRF5. The *Mc*CRF3 and *Mm*CRF3 proteins had a 96.5% similarity with each other and a 36.55% and 37.94% similarity with *Sl*CRF5, respectively (Figures 2.13, 2.14, and 2.15). ExPaSy ProtParam (Gasteigner *et al* 2005) was used to estimate the molecular weights of the protein sequences inferred from *Mc*CRF3, *Mm*CRF3, and *Sl*CRF5 sequences, with a calculated molecular weight of 32.16kDa (*Mc*CRF3) and 32.28kDa (*Mm*CRF3) versus 33.69 kDa for *SlCRF*5.

Attempts to assemble 5' UTR produced a contig ~500 bps in length (refer to Table 2.11). A region either lacking read data or containing highly repetitive sequence was encountered in each data set that precluded assembly. Although ca. 500 nucleotides were assembled upstream of the start codon, it was insufficient to identify any conserved promoter element motifs.

Expression experiments showed that *Mc*CRF3 has a low basal level of transcription but is up-regulated by oxidative stress and cytokinin treatment. Untreated shoots and roots had little to no product detectable by agarose gel electrophoresis after both PCR and qPCR (Figures 2.16 and 2.17 for qPCR gels). Strong clear banding was visible in treated shoots and roots, with the oxidative stress treatment having the most prominent band. Results of qPCR were in agreement with the band/no band AGE analysis, with exponential increase in product not being achieved for ~35 or more cycles. Samples showed great variation in quantification, with some samples not achieving exponential increase in product during amplification, even after 40 cycles. For shoot

tissue, average fold change when compared to untreated plants for the oxidative stress group was calculated to be 72.11 and 259.31 for the cytokinin treatment group. For roots, the average fold changes when compared to untreated plants were calculated to be 11.13 and 65.61 for oxidative stress and cytokinin treatment groups, respectively. Melting curve temperatures confirmed specific amplification of only the *Mc*CRF3, with samples exhibiting a specific peak melting temperature of 82.1°C (Tables 2.12 and 2.13).

Discussion

Clade III CRF genes have been found to play an important role in a variety of physiological functions, such as oxidative stress response and cytokinin response. Little research has been conducted on any gene outside of *Arabidopsis* and *Solanum lycopsericum*, so developing an understanding of the degree of conservation throughout a greater range of taxa should play an important role in developing a broader understanding of its significance. The alignment and analysis of 50 Asterid Clade III CRFs provided several novel Clade III motifs. The novel C-terminus motif was the most strongly supported motif in the MEME analysis, with an e-value of 2.2e-586 and was identified in 48 of 50 included sequences. The core of this motif was found to be [LY]D[QS]CFL[NK][DE][FY]FDFRSPSP[LI][IM]Y[ED]E. The CRF domain motif was found to be conserved with a VRI[SY]VTD[CG]DATD (e-value = 7.3e-066). Another motif that was strongly supported, and unique among the Clade III CRFs, was WDV[DN]DF[FL] (e-value =9.3e-06). This motif is typically located just upstream of the stop codon. The C-terminus motifs separate this clade from other CRF clades and are likely significant in their function. Results from experiments by Bernd (2012) suggest the 3' region plays an important role in the activation of the gene acting as a *trans*-activation site. These

conserved motifs most likely represent highly important functional domains, necessary for the function of the protein.

The assembly of *Marshallia* Clade III CRF genes will help increase our understanding of the conserved structure of Clade III CRFs. Initial assembly using DNA reads and CAP3 only provided partial contigs (~90%). This was most likely due to low read coverage. Assembly of a transcript via Trinity was much more successful, providing a full-length transcript with some up- and downstream elements. Mappings of individual DNA readsets offered very low and incomplete coverage. Species readsets offered better map quality but were still low in read depth. The assembly of a transcript via Trinity was successful, with the entire coding region and some 5' and 3' UTR elements being assembled, as well. While a full *Mc*CRF3 transcript was not assembled by Trinity, there were sufficient data to be mapped to the *Mm*CRF3 transcript to extend to the start codon and call a consensus. Mapping of M21r RNA read data provided much greater read depth. When compared to other genes, such as the tubulin genes, the read depth was quite low, ~50% for some readsets. Given the size of the *Marshallia* genome and the coverage and read depth of DNA reads, it is likely that this is a single copy gene.

Hand assembly of the 5' UTR to identify conserved promoter motifs was unsuccessful. This region featured very low coverage and an island of zero coverage was encountered in each data set. Approximately 500 nucleotides were assembled upstream of the start codon, but this was insufficient to identify the conserved promoter motifs of interest. Amplification and sequencing of this region will be needed to characterize the promoter elements due the difficulty in assembling them from these data sets.

Comparison of the *Mc*CRF3 and *Mm*CRF3 proteins with the *Solanum* Clade III CRF protein, *SlCRF5*, showed strong conservation within functional domains and motifs identified in

the Asterid Clade III CRF alignment. The CRF domains varied by 18 positions between *Sl*CRF5 and *Mc*CRF3 and only 15 with *Mm*CRF3. The core of the domain exhibited strong conservation varying by only one position (Figure 2.8). The putative 3' *trans*-activation domain was found to differ from *Sl*CRF5 in four positions in *Mc*CRF3 and three in *Mm*CRF5 (Figure 2.9). The core of the domain was highly conserved, sharing the DF[FL]DFR[SI]PSP[LI]M motif (Figure 2.1). This strongly conserved sequence of the motif likely represents the most important sequence positions for the function of this domain. The MAPk domain motif exhibited no variation between the three taxa, suggesting that is highly significant for the function of the protein (Figure 2.10). The two *Marshallia* Clade III CRF sequences featured several SNPs relative to each other, with 10 resulting in amino acids changes (Tables 2.4, 2.5, 2.9, and 2.10). While most occurred in regions that exhibit higher levels of variation between taxa, several of these occurred in conserved motifs: SNPs 107 and 141 are located within the CRF domain and 709, 711, and 720 are located within the novel 3' motif. This suggests that there is some plasticity in sequences of closely related taxa while conserving function.

Expressions experiments showed that this gene has a low basal level of transcription and is up-regulated by both oxidative stress and cytokinin treatment. *Mc*CRF3 PCR and qPCR products of untreated plants exhibited no banding when analyzed by agarose gel electrophoresis (Figures 2.15 and 2.16). Strong clear banding was exhibited by treatment groups, with the oxidative stress group exhibiting the strongest band. Results of qPCR were in agreement with the band/no band AGE analysis, with exponential increase in product not being achieved for ~37 or more cycles. Samples were of sufficiently low quantity to preclude calculating fold change for every sample as many samples were unable to be consistently detected. Average fold changes were calculated for samples achieving exponential increase during amplification and showed that

*Mc*CRF3 was up-regulated by both oxidative stress and cytokinin treatment in both shoot and root tissue. The highest fold change in expression was found to be in the shoot tissue due to cytokinin treatment. These results are in agreement with previous work on Clade III CRFs, which showed strongest up-regulation in aerial organs by cytokinin treatment (Zwack *et al* 2012) and are also consistent the read depth provided by three *Marshallia* RNA datasets (M2.9, M3.9, and M20r), which were not sufficient to fully assemble a transcript, possible suggesting low expression at time of extraction. The results of this analysis show that there is conservation in *Marshallia* for, at least, some regulatory mechanisms found in *AtCRF5* and *SlCRF5*.

Using NGS data sets of DNA that were designed for use in genome-skimming approaches proved to be difficult for the *Marshallia* Clade III CRF as read depth was insufficient for full assembly. The data were very useful in preliminary assemblies and in designing primers. RNA datasets were more useful, as a full transcript was assembled from accession M21r. Overall, use of NGS technology in this context can be useful and informative, though more traditional techniques (*e.g.*, PCR and capillary sequencing) may be needed for supplementation.

Literature Cited

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

2. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D. L. 2005. GenBank. *Nucleic Acids Research, 33*(Database issue), D34–D38. doi:10.1093/nar/gki063

3. Cutcliffe, J.W., Hellmann, E., Heyl, A. and Rashotte, A.M. 2011. CRFs form protein–protein interactions among each other and with members of the cytokinin-signaling pathway in Arabidopsis via the CRF domain. *J. Exp. Bot.* **62**, 4995–5002.

4. Gupta, S. 2013. Characterization of CRF domain containing ERF genes- *Solanum lycopersicum* Cytokinin Response Factors *Sl*CRF3 and *Sl*CRF5 in tomato development.

5. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. 2005. *Protein Identification and Analysis Tools on the ExPASy Server;* (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press. pp. 571-607.

6. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol*. **29**, 644-52.

7. Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res*., **9**, 868-877.

8. Ketelsen, B. 2012. Characterization of a Cytokinin Response Factor in Arabidopsis thaliana.

9. Mangani, E., Sjolander, K., and Hake, S. 2004. From endonucleases to transcription factors: evolution of the AP2 DNA binding domain in plants**.** *Plant Cell*, **16**, 2265-2277.

10. Miller, C.O., Skoog, F., Okomura, F.S., von Saltza, M.H., and Strong, F.M. 1956. Isolation, structure and synthesis ofkinetin, a substrance promoting cell division. *J. Am. Chem. Soc*. **78**, 1345–1350.

11. Miller, C.O., Skoog, F., von Saltza, M.H., and Strong, F. 1955. Kinetin, a cell division factor from deoxyribonucleic acid. *J. Am. Chem. Soc*. **77**, 1392.

12. Milne, I., Stephen, G., Bayer, M., Cock, P.J.A., Pritchard, L., Cardle, L., Shaw, P.D. and Marshall, D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* **14**, 193-202.

13. Müller, B., and Sheen, J, Cytokinin signaling pathway. *Sci. STKE* (Connections Map, as seen October 2007)

14. Okamuro, J., Caster, B., Villarroel, R., van Montagu, M., and Jofuku, K. 1997. The AP2 domain of *APETALA2* defines a large new family of DNA binding proteins in *Arabidopsis***.** *Proc Natl Acad Sci USA.* **94**, 7076-7081.

15. Okazaki, K., Kabeya, Y., Suzuki, K., Mori, T., Ichikawa, T., Matsui, M., Nakanishi, H., and Miyagishimaa, S. 2009. The PLASTID DIVISION1 and 2 components of the chloroplast division machinery determine the rate of chloroplast division in land plant cell differentiation. *Plant Cell*. **21**, 1769–1780.

16. Rashotte, A.M., Mason, M.G., Hutchinson, C.E., Ferreira, F.J., Schaller, G.E., Kieber, J.J. 2006. A subset of Arabidopsis AP2 transcription factors mediates cytokinin responses in concert with a two-component pathway. *PNAS* **103**, 11081–11085.

17. Rashotte, A.M. and Goertzen, L.R. 2010. The CRF Domain defines Cytokinin Response Factor proteins in plants. *BMC Plant Biology* **10**:74.

18. Rodriguez, M.C., Petersen, M., and Mundy, J. 2010. Mitogen-activated protein kinase signaling in plants. *Annu Rev Plant Biol* **61**: 621–49. doi:10.1146/annurev-arplant-042809-112252. PMID 20441529.

19. Schmülling, T., Schäfer, S., and Romanov, G. 1997. Cytokinins as regulators of gene expression. *Physiologia Plantarum*, **100**(3), 505–519. doi:10.1111/j.1399-3054.1997.tb03055.x

20. Shi, X., Gupta, S. and Rashotte, A. M. 2012. *Solanum lycopersicum cytokinin response factor* (*SlCRF*) genes: characterization of CRF domain-containing ERF genes in tomato. *Journal of Experimental Botany* **63**: 973–982.

21. Skoog, F., and Miller, C.O. 1957. Chemical regulation of growth and organ formation in plant tissues cultured in vitro. *Symp Soc Exp Biol*. **11**, 118-30.

22. Strnad, M. 1997. The Aromatic Cytokinins. *Physiol. Plant*. **101**, 674–688.

23. Swofford, D. 1998. PAUP 4.0: phylogenetic analysis using parsimony. Smithsonian Institution.

24. Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

25. Zwack, P.J., Shi, X., Robinson, B.R., Gupta, S., Compton, M.A., Gerken, D.M., Goertzen, L.R., and Rashotte, A.M. 2012. Vascular Expression and C-Terminal Sequence Divergence of Cytokinin Response Factors in Flowering Plants. *Plant Cell Physiol.* **53**, 1683–1695.

26. Zwack, P.J., Robinson, B.R., Risley, M.G. and Rashotte, A.M. 2013. Cytokinin Response Factor 6 Negatively Regulates Leaf Senescence and is Induced in Response to Cytokinin and Numerous Abiotic Stresses. *Plant Cell Physiol.* **54**, 971–981.

Tables and Figures

Table 2.1. Protocol for PCR of *Mc*CRF3.

Stoichiometry		PCR Parameters	
G-Bio Taq	10 µL	T_m	94.0C 30s
Primer	0.4 µL	T_a	55.0C 30s
Template	9.6 µL	T_e	72.0C 30s
		Cycles	35

Table 2.2. Protocol for qPCR of *Mc*CRF3.

Stoichiometry		qPCR Parameters	
Taq-SYBR	10 µL	T_m	95.0C 20s
Primer Mix	0.4 µL	T_a	55.0C 20s
Template	9.6 µL	T_e	68.0C 30s
		Cycles	40
		Melting Curve	95.0C 15s
			60.0C 15s
			20m
			95.0C 15s

Table 2.3. Primers used in PCR reactions for expressions experiments.

Primer	Sequence
Clade III CRF Forward	GCTTCTGGTTCTGTGTCCGA
Clade III CRF Reverse	CCAAACCGTAACACGGAGGT
EF1A Forward	GATGATTCCCACCAAGCCCA
EF1A Reverse	CAAACAACCGACGAACCCAC

Table 2.4. Sequence of the *Marshallia caespitosa* Clade III CRF gene coding region. SNPs between the *Mc*CRF3 and *Mm*CRF3 coding regions are highlighted in yellow.

*Mc*CRF3	ATGAAACTAGACTTCATGGGATCTTCTCCTAAATTTAGAGTCAATCTC
	ACCGTCACCATCAAACAATCGGAACTCGATTCACCAAAAACGGTAAC
	GATTTCCATGAACGATCGCGATGCTACGGACTCTTCAAGTGACGAAG
	ACCACAACGAATTGGGTCACCGAAAAATTAAAAGGTATGTAAATGTG
	ATTCAGTTCGAAGACAACTGTTGTGGGAGAAATTTGAGTGGAAGTGA
	TGGTAGTGATAAAGGGAAGAAAAAACAGAGTCGCCGGATGAAAGAA
	CCAGTAAGTTCGGGAACCGAACGGAAGTTTAGGGGAGTAAGGCGACG
	GCCATGGGGAAGGTGGGCGGCGGAGATTCGTGATATGGGGGTGAGGG
	TATGGTTGGGTACATACGATACGGCGGAGGAGGCGGCATTAGCTTAT
	GATCGGAGAGCGATTGAACTACATGGATGGAAAGCTCAGACGAACTT
	TTTGCAACCACCGCGTTCAGAAGTTGCAGTACCGGTGATAGCTTCTGG
	TTCTGTGTCCGATCAGTGTTCCGGGAAGGAGTTGCGGGGTGTTTCCTC
	GCCGACCTCCGTGTTACGGTTTGGTAAAACAGAGGCGGAGTCTGAGA
	AGCTGGATGAACAGAAGCAAAGTGAATCAAATGTCAGAGATGATGAT
	TTCGGGTATGATTGGGATTTGGAATACGATTTCTTAGACTTTCGGATA
	GCGTCTCCGATGATGGTGGAAGAGATTGATTTGGGAAGAAGAATGAT
	GTGGGAGGTAGAGGACGATATGAAACCAAGAGTGTGGGATGTGGATG
	GTTGCTTCCAAGATCCTGTGATTGGAGAGTGGTTAGATGATTAA

Table 2.5. Sequence of the *Marshallia mohrii* Clade III CRF gene coding region. SNPs between the *Mc*CRF3 and *Mm*CRF3 coding regions are highlighted in yellow.

*Mm*CRF3	ATGAAACTAGACTTCATGGGATCTTCTCCTAAATTTAGAGTCAATCTC
	ACCGTCACCACCAAACAATCGGAACTCGATTCACCGAAAACGGTAAC
	GATTTCCATGACCGATCGCGATGCTACGGACTCTTCAAGTGACGACGA
	CCACAACGAATTGGGTCACCGGAAAATTAAAAGGTATGTAAATGTGA
	TTCAGTTCGAAGACAACTGTTGTGGGAGAAATTTGAGTGGAAGTGAT
	GGTAGTGATAAAGGGAAGAAAAAACAGAGTCGCCGGATGAAAGAAC
	CAGTAAGTTCGGGAACCGAACGGAAGTTTAGGGGAGTAAGGCGACGG
	CCGTGGGGAAGGTGGGCGGCGGAGATTCGTGATATGGGGGTGAGGGT
	ATGGTTGGGTACATACGATACCGCGGAGGAGGCGGCATTAGCTTATG
	ATCGGAGAGCGATTGAACTACATGGATGGAAAGCTCAGACGAACTTT
	TTGCAACCACCGCGTTCAGAAGTTGCAGTACCGGTGATAGCTTCTGGT
	TCTGTGTCCGATCAGTGTTCCGGGAAGGAGTTGCGGGGTGTTTCCTCG
	CCGACCTCCGTGTTACGGTTTGGTAAAACAGAGGCGGAGTCTGAGAA
	GCTGGATGAACAGAAGCAAAGTGAATCAAATGGCAGAGATGATGATT
	TCGGGTATGATTGGGATTTGGAATACGATTTCTTAGACTTTCGGATAC
	CATCTCCGATAATGGTTGAAGAGATTGATTTGGGAAGAGGAATGATG
	TGGGAGGTAGAGGACGATATGAAATCAAGAGTGTGGGATGTGGATGG
	TTGCTTCCAAGATTCTGTGGTTGGAGAGTGGTTAGATGATTAA

Table 2.6. Mapping data for species DNA data sets.

Species	N	% Cvg.	Avg. Cvg.	Max Cvg.	# reads	% mismatch
M. caespitosa	2	80.30	2.38	14	31	5.8
M. graminifolia	3	89.30	2.35	7	30	4.4
M. grandiflora	2	89.64	2.97	7	31	3.4
M. legrandii	1	89.58	2.16	7	27	4.3
M. mohrii	2	97.82	3.28	8	40	3.9
M. obovata	3	85.42	2.57	11	38	3.8
M. ramosa	2	80.40	2.01	9	25	4.1
M. trinervia	2	87.78	3.01	8	39	3.5

Table 2.7. Clade III CRF mapping data for arbitrarily sampled individual data sets representing each species of *Marshallia*.

Accession	% Cvg	Avg Cvg	Max Cvg	# reads	% mismatch
M2	49.62	0.66	3	9	5.7
M3	31.91	0.45	3	9	3.8
M5	72.54	0.90	2	11	6.1
M17	78.79	1.27	5	14	3.1
M19	74.62	1.52	6	18	3.3
M24	90.06	2.13	7	25	3.9
M26	44.60	0.78	5	10	5.6
M31	89.58	2.16	7	27	4.3

Table 2.8. Mapping data for arbitrarily sampled species data sets mapped to other nuclear encoded genes.

Gene	Species	% Cvg	Avg Cvg	Max Cvg	# reads	% mismatch
Actin	*M. grandiflora*	98.46	7.65	21	156	5.3
	M. obovata	100	4.84	21	106	5.9
α-tubulin	*M. ramosa*	99.99	2.30	6	47	2.5
EF1A	*M. grandiflora*	99.09	7.72	15	156	4.4
π-tubulin	*M. caespitosa*	88.52	1.88	6	47	1.2
	Mgraminifolia	91.62	3.40	9	82	1.5

Table 2.9. Amino acid sequence for the *Mc*CRF3 protein. Amino acid differences between *Mc*CRF3 and *Mm*CRF3 are highlighted in yellow.

*Mc*CRF3	MKLDFMGSSPKFRVNLTVTIKQSELDSPKTVTISMNDRDATDSSSDEDHN
	ELGHRKIKRYVNVIQFEDNCCGRNLSGSDGSDKGKKKQSRRMKEPVSSGT
	ERKFRGVRRRPWGRWAAEIRDMGVRVWLGTYDTAEEAALAYDRRAIEL
	HGWKAQTNFLQPPRSEVAVPVIASGSVSDQCSGKELRGVSSPTSVLRFGK
	TEAESEKLDEQKQSESNVRDDDFGYDWDLEYDFLDFRIASPMMVEEIDLG
	RRMMWEVEDDMKPRVWDVDGCFQDPVIGEWLDD

Table 2.10. Amino acid sequence for the *Mm*CRF3 protein. Amino acid differences between *Mc*CRF3 and *Mm*CRF3 are highlighted in yellow.

*Mm*CRF3	MKLDFMGSSPKFRVNLTVTTKQSELDSPKTVTISMTDRDATDSSSDDDHN
	ELGHRKIKRYVNVIQFEDNCCGRNLSGSDGSDKGKKKQSRRMKEPVSSGT
	ERKFRGVRRRPWGRWAAEIRDMGVRVWLGTYDTAEEAALAYDRRAIEL
	HGWKAQTNFLQPPRSEVAVPVIASGSVSDQCSGKELRGVSSPTSVLRFGK
	TEAESEKLDEQKQSESNGRDDDFGYDWDLEYDFLDFRIPSPIMVEEIDLGR
	GMMWEVEDDMKSRVWDVDGCFQDSVVGEWLDD

Table 2.11. Sequence of 5' UTR assembled from all *Marshallia* spp. read sets. This region extended approximately 500 bps upstream of the start codon.

GGAGAGAGAGAGGTCTCAATGTGGGAGTAGCACAAGGTCTTATTTTTTCAATGTGG
GATGGGGCTATTTCAATATTAAAAAAAGTAAATGTGTCTAGATGTTGGAGAGAGAG
GGGATTAAGAAACACCTTATGTGCTACATTTGAAGCAGCATGTAAACGTACATTATG
TAGTTCTTACCGTTCTTAACACAAAATGACTTCTTGCATAAGCGTTCCTCTATACATA
TATATAGGGTGCNGTTCATGCGAGAACCAAGTTTATGCCGAGAATCACGAGAACNA
TTGGTATTATTGTAATTAATTACAAATTACATAAAATATATTTTCGGTTCAACCAAAA
ACTAATCGACATATATGCATCCAACGAACGTAACCAAATCAAANGATCTTTCCTACG
ATCTCTTCTTCATCTTTAGCTTCCACCGTCGTACACATCTTTTCCGGCGATCAATTACT
TCTATTCTATTT

Table 2.12. qPCR quantitative data from shoot analysis.

Sample	Mean ΔCt	Mean ΔΔCt	Avg. F.C.
Control	10.71		
Oxidative	5.57	-5.13	72.11
Cytokinin	4.23	-6.54	259.31

Table 2.13. qPCR quantitative data from roots analysis.

Sample	Mean ΔCt	Mean ΔΔCt	Avg. F.C.
Control	3.36		
Oxidative	-0.12	-3.48	11.13
Cytokinin	-2.68	-6.04	65.61

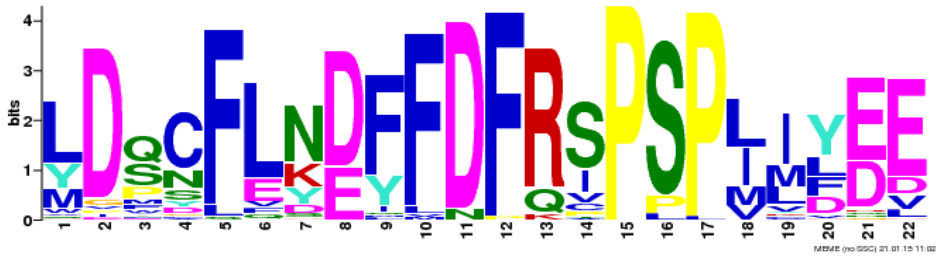

Figure 2.1. MEME analysis output for the putative 3' *trans*-activation domain motif. This motif occurs downstream of the MAPk domain and was supported by an e-value of 2.2e-586.

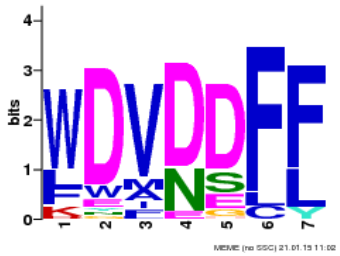

Figure 2.2. MEME analysis output for the putative 3' *trans*-activation domain motif. This motif typically occurs just upstream of the stop codon and was supported by an e-value of 9.3e-0.6.

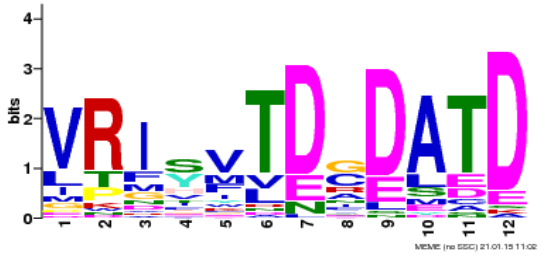

Figure 2.3. The CRF domain motif. This motif is the first of the motifs identified by the MEME analysis to occur in the amino acid sequence. It is supported by an e-value of 7.3e-066.

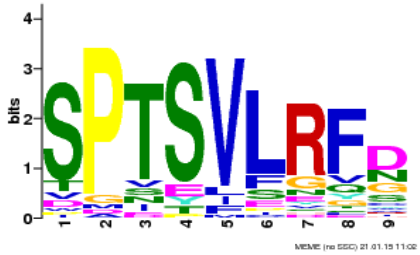

Figure 2.4. MEME analysis output for the putative MAPk domain motif. This motif occurs downstream of the AP2/ERF domain and was supported by an e-value of 6.1e-054.

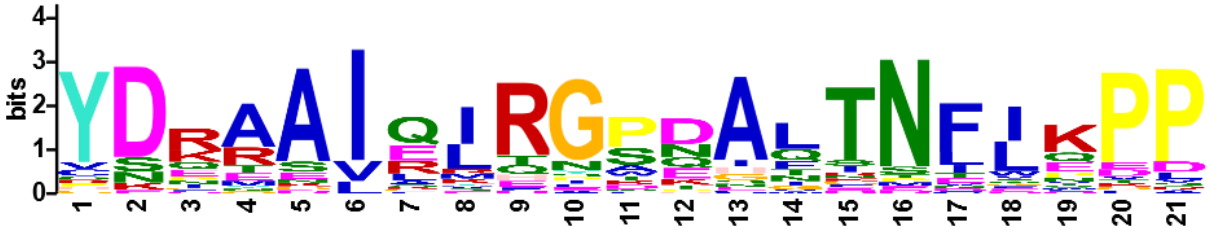

Figure 2.5. MEME output for the most statistically supported AP2/ERF domain motif. This motif is the last of the identified in this analysis to occur in the AP2/ERF domain and is supported by an e-value of 5.4e-221.

Figure 2.6. MEME output for the second most statically supported AP2/ERF domain motifs. This is the second AP2/ERF domain motif identified in this analysis to occur in the AP2/ERF domain and is supported by an e-value of 5.2e-204.

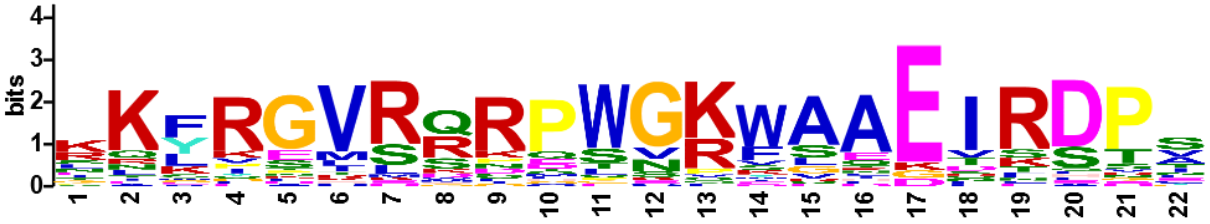

Figure 2.7. MEME output for the third most statically supported AP2/ERF domain motifs. This is the first AP2/ERF domain motif identified in this analysis to occur in the AP2/ERF domain and is supported by an e-value of 6.5e-200.

Figure 2.8. Relative position map of conserved domains/motifs, highlighted in red, of the *Marshallia* Clade III CRF proteins.

Figure 2.9. Comparison of the CRF domains of *SlCRF5* (top), *Mc*CRF3, and *Mm*CRF3 (bottom) Clade III CRF proteins. This motif occurred in amino acid positions 26 - 47 in *Sl*CRF5 and 27-48 in *Mc*CRF3 and *Mm*CRF3.

Figure 2.10. Comparison of strongly conserved motif in putative *trans*-activation domain in *SlCRF5* (top), *Mc*CRF3, and *Mm*CRF3 Clade III CRF proteins. This motif occurred in amino acid positions 237 - 253 in *Sl*CRF5 and 230 - 247 in *Mc*CRF3 and *Mm*CRF3.

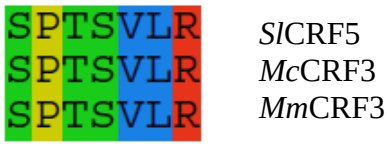

Figure 2.11. Comparison of the putative MAPk domain motif in *SlCRF5* (top), *Mc*CRF3, and *Mm*CRF3 (bottom) Clade III CRF proteins. This motif occurred in amino acid positions 180 - 186 in *Sl*CRF5 and 188 - 195 in *Mc*CRF3 and *Mm*CRF3.

Figure 2.12. Comparison of 3' motif of *trans*-activation domain in *SlCRF5* (top), *Mc*CRF3, and *Mm*CRF3 Clade III CRF proteins. This motif occurred in amino acid positions 283 - 287 in *Sl*CRF5 and 277 - 281 in *Mc*CRF3 and *Mm*CRF3.

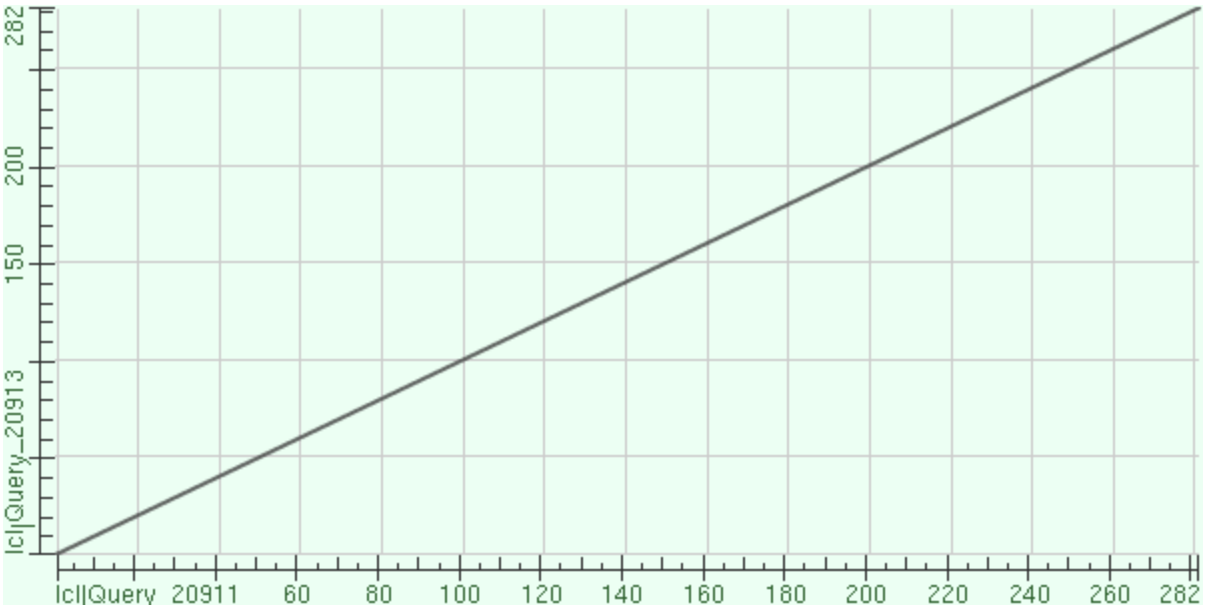

Figure 2.13. BLAST output for the comparison of the two newly described *Marshallia* Clade III CRF proteins. Percent identity of *Mc*CRF3 and *Mm*CRF3 was calculated 96.5 by BLASTn.

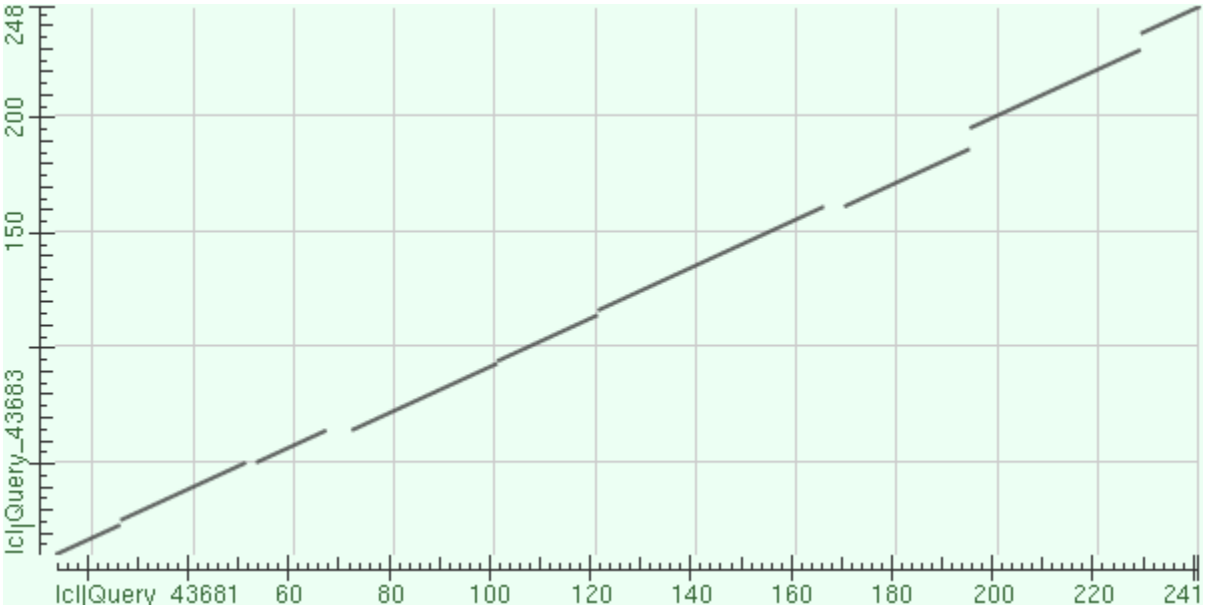

Figure 2.14. Comparison of *Mc*CRF3 to *Sl*CRF5. BLAST output. Percent identity of *Mc*CRF3 and *Sl*CRF5 was calculated 36.55 by BLASTn.

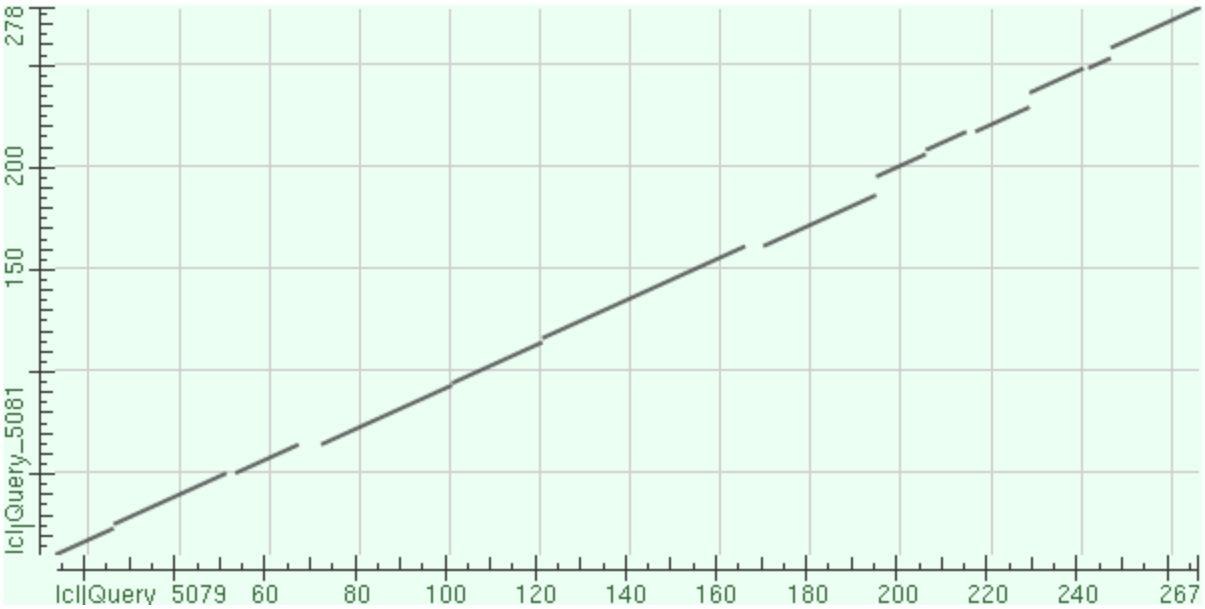

Figure 2.15. BLAST output for comparison of *Mm*CRF3 to *Sl*CRF5. Percent identity of *Mm*CRF3 and *Sl*CRF5 was calculated 37.94 by BLASTn.

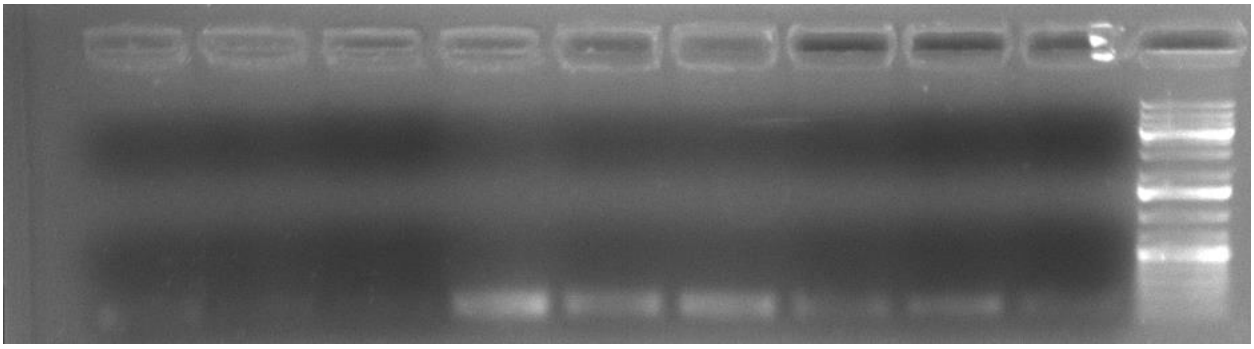

Figure 2.16. Gel visualization of qPCR products for shoots. Controls = 1-3, Oxidative stress = 4-6, Cytokinin = 7-9. No banding was observed for controls while bands were observed for treatments, suggesting up-regulation. PCR amplicons were approximately 200 bps in size.

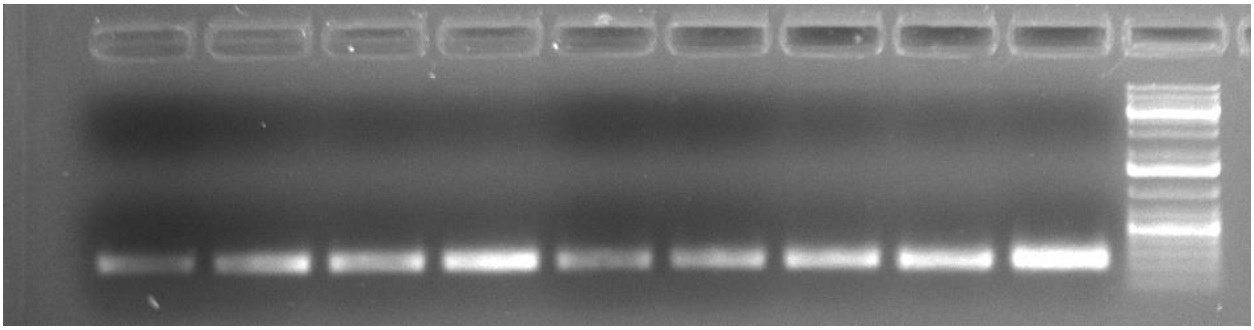

Figure 2.17. EF1A qPCR products for shoots. All samples exhibited banding. Order of samples is as above. PCR amplicons were approximately 200 bps in size.

Chapter Three. A Survey of RNA Editing in the *Marshallia* Plastid Genome Utilizing NGS Technology

Abstract

RNA editing is an RNA maturation process in which one nucleotide is modified into or replaced with another. Within angiosperm plastid genomes, ~35 C-U and one A-I edits have been identified, though, RNA editing can be a highly variable process, with species possessing unique edit sites. To identify plastid RNA editing within the plastid genomes of multiple *Marshallia* species, DNA and RNA data were generated using ILLUMINA NGS platforms and aligned for direct comparison. Thirty-eight editing sites were identified (31 in coding regions), with 24 being shared among other taxa. Partial editing was also quantified and revealed that individuals of *M. mohrii* (accessions M20r and M21r), a putative allopolyploid, exhibited less efficient editing (as determined by the frequency of the edited base) than that found in the related diploid species included in this study (*M. trinervia* and *M. obovata*; ~78% of edit sites exhibiting partial editing in *M. mohrii* vs. 25-30% in diploid species). The most common effects of editing were alteration of codons in the 2nd position altering the amino acid residue to a Leucine.

Introduction

RNA editing is a post-transcriptional process in which one nucleotide is substituted for or modified into another. This can include cytidine-to-uridine deamination in mRNAs and tRNAs, uridine to cytidine amination in mRNAs and adenosine-to-inosine deamination in tRNAs (Chateigner-Boutin and Small 2010, Schuster *et al* 1990). These changes often result in no change or the conservation of translated amino acids in the protein sequence, although editing events have also been identified that create diversification of sequence (Inada *et al* 2004). Additionally, these RNA edits can also create necessary stop codons and create or silence stop codons (Bock *et al* 1994, Wolf *et al* 2004, Kugita *et al* 2003, Jobson and Qiu 2008). These edits have also been found to occur in UTRs of the RNAs and in the mitochondrial and plastid genomes though no editing has been found in ribosomal RNAs (Chateigner-Boutin and Small, Bock, Takenaka 2008, Mulligan *et al* 1999, Tangphatsornruang *et al* 2011). RNA editing can also occur in only a fraction of molecules of a given transcript, a phenomenon known as partial editing. Partial editing has been thought to function as a transcription regulating mechanism as only edited transcripts will undergo full maturation and function properly (Miyata *et* al 2002, Miyata and Sugita 2004). Methods for quantifying partial editing include comparing genomic DNA and cDNA PCR amplicons, NGS and computational approaches, and High Resolution Melting (HRM) of PCR amplicons (Miyata and Sugita 2004, Chateigner-Boutin and Small 2007).

A-I editing of the tRNA-ARG (ACG) has been found in all major lineages investigated, from bacteria to angiosperms (Chateigner-Boutin and Small 2010, Delannoy *et al* 2009). This suggests that the A-I editing site and mechanism evolved early, most likely in the bacterial ancestor of green plant chloroplasts. Plastid C-U RNA editing is found in all land plant lineages

except the marchantiids (Freyer *et al* 1997). Plastid C-U RNA editing has not been found in green algae, suggesting that its evolution coincided with the evolution of land plants. Several hypotheses for the coincidence of these major evolutionary events have been proposed. One hypothesis is that RNA editing evolved and was conserved as a mechanism to protect and conserve DNA, RNA and amino acid sequences from UV radiation, correcting C-to-T mutations (Maier *et al* 2008).

Approximately 435 edit sites have been identified in angiosperm organellar genomes (~400 in the mitochondrial and ~35 in the chloroplast). RNA editing can be highly variable, even between closely related taxa. Some taxa, such as those in the flowering plant family Geraniaceae, can have greatly reduced editing (Takenaka 2013, Chateigner-Boutin and Small 2010, Zhang *et al* 2013, Nugent and Palmer 1991, Bock 1998, Takenaka 2008, Dombrovska and Qiu 2004, Duffy *et al* 2009, Sper-Whitis *et al* 1996). RNA editing can even vary within an individual as it has been found to be tissue and life-stage specific (Miyata *et al* 2002, Miyata and Sugita 2004). While there is no known cause of the great discrepancy in edit site quantity between organelles or between various taxa, it is thought that the number of edit sites may be influenced by the number of editing proteins of the PPR protein family present, known as PPR-DYW proteins. Angiosperms have been found to have around 100 PPR-DYW genes while *P. patens*, a moss, only has ten. This moss not only has fewer PPR-DYW proteins, but also fewer editing sites (12 total; 1 plastid, 11 mitochondrial) (Tasaki and Sugita 2010).

Chloroplast RNA editing has been found to consist only of approximately 35 C-U deamination sites (Tsudzuki *et al* 2001) and one A-I deamination site. The 35 C-U edits can be found in approximately ~20 genes and several non-coding regions. The single A-I deamination has been found to occur in the tRNA trnR-ACG. Some C-U edits have been found to be

conserved across a range of taxa including *Nicotiana*, *Arabidopsis*, *Oryza*, and *Pisum sativum.* The number of editing sites has been found to vary across these organisms, with 34 C-U edits in *Arabidopsis*, 31 in *Nicotiana*, and 27 in *Pisum sativum* (Inada *et al* 2004, Chateigner-Boutin and Small 2007). Three genes (*rpl2, ndhD,* and *psbL*) have been found to contain conserved C-U edits within the start codon that are required for translation (Chaudhuri and Maliga 1996, Inada *et al* 2004).

RNA editing has been found to require a *cis*-element for edit site identification and a *trans*-element editing protein for the reaction. *cis*-elements have been found to occur in a wide range of positions, from 1 to ~50 bps upstream and even up to ~25 bps downstream of the edit site, though no common identification motif has been identified (Giege and Brennicke 1994). Often, the -1 bp position relative to the first position of the start codon has the greatest effect on edit site identification (Hermann and Bock 2000). Some evidence suggests that there may common nucleotide triplets preceding the edit site that could play a role in identification, with certain triplets never occurring before an edit site (Du *et al* 2007).

Proteins thought to perform the editing are typically members of the PPR (pentatricopeptide repeat) containing family of proteins. This is a broad family of proteins, with a variety of subgroups, varying in their motif content. These proteins have been found to contain a large N-terminus region that contains the standard PPR motif of 35 amino acids and degenerate repeats that have no known function. P-class PPR proteins contain only the 35 amino acid PPR motif. PLS-class PPRs contain a PPR (P) motif, a long (L) motif, and a short (S) motif. The PPR-DYW proteins are a member of the PLS-class and contain the P motif as well as a DYW domain, named after the aspartate-tyrosine-tryptophan motif that is thought to confer their enzymatic abilities, and an extended (E) motif (Chateigner-Boutin and Small 2010, Okuda *et al* 2009).

Although there are many edit sites in plant organellar genomes, only a few RNA editing proteins have been identified in *Arabidopsis*. One such protein, AtTadA (tRNA adenosine deaminase arginine), has been found to function in adenosine deamination (Karcher and Bock 2009). Though the enzyme has been identified for A-I edits in the chloroplast, the plastid C-U editing enzymes are still largely unknown. AtECB2 is a PPR-DYW protein that is required for editing of the accD editing site in *Arabidopsis.* This protein contains a conserved $HxEx_nCxxC$ motif, similar to the activated site of cytidine deaminase (Yu *et al* 2009). OTP82 is another PPR-DYW protein that has been found to function in the editing of the plastid genes ndhB and ndhG, though there is evidence that the DYW motif present in this gene is not required for editing (Okuda *et al* 2010). PpPPR_71 is a PPR-DYW protein that has been shown to bind with *ccmFc* mitochondrial transcripts to perform editing (Tasaka and Sugita 2010). One editing protein is CRR4, a protein that does not contain a PPR domain and works in conjunction with DYW1, another editing protein that lacks a PPR domain, to edit *ndhD* edit site 1 in *Arabidopsis thaliana* (Boussardon *et al* 2012). While many of the molecular machinery remains to be identified, experiments have shown that the location of these *trans*-factors that function in editing must be present in the nucleus (Bock and Koop 1997).

The *trans*-factors that function in editing are often site-specific, with the *cis*-element of a given edit site being recognized by a specific PPR protein of the E or DYW groups (Bock and Koop 1997, Miyamoto *et al* 2002), though some proteins have been found to be more general in site recognition (Choury and Araya 2006). PPR motifs contain two alpha helices that interact with the *cis*-elements, with specificity being determined by discrimination between purines and pyrimidines. There are four models proposed to explain the deamination reaction: 1) deamination is catalyzed by the E domain, 2) the DYW motif catalyzes the deamination in *cis*, 3) the DYW

catalyzes the deamination in *trans*, and 4) the deamination could be catalyzed by an unknown enzyme recruited by the E domain (Salone *et al* 2007, Chateigner-Boutin and Small 2010).

Several edit site predicting software have been produced. Prep-Mt, Prep-Cp, Prep-Aln and CURE both identify edit sites within organellar genomes. These methods use the knowledge of editing (e.g., typically create more hydrophobic and phylogenetically conserved amino acids) and DNA sequence data. These methods can be very effective but would not detect edits that create a residue that is phylogenetically divergent from the Marchantiids. Other algorithms are being developed that do not use phylogenetic information for site detection. Du, He and Li (2007) developed an algorithm based a SVM algorithm with a triplet scoring system. Although these programs can have high success rates, the only methods to truly identify edit sites are through experimentation comparison of DNA sequence data to RNA transcript sequence data (Mower 2005, Mower 2009, Du and Li, Du He and Li).

Most research on RNA editing has been conducted on *Arabidopsis*, *Nicotiana*, *Oryza*, and *Pisum sativum* by comparing PCR amplicons of genomic DNA and cDNA. Identification of RNA edit sites has also come as a byproduct of genome assembly using both whole genomic DNA and ESTs or Transcriptomic data (Timme *et al* 2007). Research elucidating RNA editing sites in non-model taxa is important for increasing our understanding of the evolution and phylogenetic distribution of RNA editing sites. Also, understanding where and how transcripts are edited is important for understanding how the genes and their protein products function. For phylogenetic analyses, using unedited nucleotide or amino acid sequences can lead to errors as conservation in the protein will not be shown (Duffy *et al* 2009). Prior to the advent of NGS technologies, identification of editing sites was time and labor intensive. It required the amplification and sanger-sequencing of plastid genes and transcripts and utilized less efficient

assembly methods. This project aims to identify edits sites within the chloroplast genome of *Marshallia* spp. using NGS technology. The use of ILLLUMINA sequence data and bioinformatic tools will allow for a large-scale and streamlined analysis. The information gained from this research will help increase understanding of the phylogenetic distribution of plastid RNA editing sites and increase our understanding of plastid genetics within *Marshallia.* Because of the few studies that have investigated RNA editing in closely related groups, especially those that feature hybrids, this study will illuminate potentials patterns in the inheritance of edit sites from parent species and how editing could be conserved or divergent in closely related taxa. Few, if any, differences are expected to be identified in the plastid genome due to their high degree of similarity (Hansen and Goertzen, UNPUBLISHED). Approximately 35 C-U edit sites are expected to be identified, occurring primarily within coding regions.

Methods

Nucleic acid extractions were performed on 4 individuals of *M. mohrii* (accessions M21.1 and M24.1 for DNA; M20r and M21r for RNA), 4 individuals of *M. obovata* (accessions M3.1, M4.1, and M32.2 for DNA; M3.9 for RNA) and 3 of *M. trinervia* (accessions M2.3 and M33.6 for DNA; M2.9 for RNA). DNA were extracted from fresh leaf material using a modified CTAB protocol described by Doyle and Doyle (1987) or an E.Z.N.A.® Plant DNA extraction kits (Omega Bio-tek, Inc., Norcross, GA) per manufacturer protocol. RNA were extracted from fresh leaf material using Plant RNA Extraction kits (Qiagen, Hilden, Germany) per manufacturer protocol. DNA samples were submitted to the HudsonAlpha Institute for Biotechnology (Huntsville, AL) for paired-end library prep and 100bp sequencing via an ILLUMINA (ILLUMINA Inc., San Diego, CA) HiSEQ2000 platform. RNA samples were submitted to the Auburn University Genomics and Sequencing Laboratory (Auburn, AL) where cDNA libraries

were prepared using an Illumina mRNA TruSeq kit, which uses mRNA beads with poly-T sequences to isolate mRNA. Sequencing was performed using an ILLUMINA HiSEQ1500 platform.

Whole genomic DNA reads were cleaned using Sickle (Joshi and Fass 2011) to a quality score of 20 and minimum length of 85. Cleaned reads of M3.1 were then assembled *de novo* via Ray v. 2.3.1 (Boisert *et al* 2010). Assembled chloroplast contigs were then identified via BLASTn using *Guizotia abyssinica*, a fellow member of the Heliantheae tribe of Asteraceae, chloroplast genome (NCBI GenBank accession EU549769.1) as the query and extracted via a pearl script. Chloroplast contigs were manually assembled in SeaView v. 4.4.2 (Gout *et al* 2010). Reads were mapped via Bowtie2 v. 2.1.0 (Langmead and Salzberg 2012) to the assembled chloroplast to assess the accuracy of assembly. Mappings were visualized in Tablet v. 1.14.04.10 (Milne *et al* 2013). The M3.1 chloroplast genome then served as the reference genome in future assemblies. After the assessment of the reference chloroplast genome, cleaned DNA reads of each were mapped via Bowtie2 to the reference and a strict consensus sequence called via an Eric Archer .plp consensus script in R. RNA reads were mapped via Bowtie2 and a consensus sequence was called. The consensus calling criteria for RNA were set so that an ambiguity symbol would be called for any positions for which less than 80% of the bases mapped to that position were of one base. Proportion of bases per position as calculated by the consensus calling script was used in determining the extent of editing per site. DOGMA (Wymen *et al* 2004) was used to annotate both plastid genome and transcriptome. The chloroplast genome and transcriptomes were then aligned in MAFFT v6.935b (Katoh 2013).

SeaView was used to visualize the alignment and to manually scan for disagreements between DNAs and RNAs. An initial round of edit site identification was performed by manually

assessing each position within the alignment. Each disagreement between DNA and RNA was assessed for its conversion type, agreement among samples, percent bases, percent error, and location. For initial consideration, a putative edit must occur across all species being represented. If a disagreement was atypical (i.e., not C-U) or was due to polymorphisms, the position was excluded. For depth and percent bases, a given position must have >20% of the edited base for consideration. For position error, no position was considered if there was more than 5% non-edited or edited bases present (i.e., not a T or C). For genes that have known edit sites in other species, but did not have edit sites identified in the initial genome scanning, a second round of identification was performed using gene and transcript sequences. Individual genes and their transcripts were aligned manually in SeaView and compared. Amino acid sequences were inferred in SeaView from the DNA and RNA sequences to determine how edits affect their sequence.

Results

In total, 31 edit sites were identified within 14 genes (Table 3.1) and seven edit sites were identified within non-coding regions (two in ndhC-ndhK , one in trnQ-rps16, one in rps4-trnS-GCU , two in psal-ycf4, and one in psbE-petL). Two of the edit sites identified were found within start codons, resulting in an ACG-AUG change. All but one edit within coding regions were found to alter the amino acid sequence (See Figures 3.1-3.3 for example alignments). One edit (petB edit 1) was found to be a silent edit, altering only the nucleotide sequence. No edit was found within the trnR (ACG) region, a gene with known A-I edits. No edits were found to be needed to cause stop codons in rps3 or psbC, two genes known to require editing to a stop codon in *Helianthus* and *Lactuca* (Figure 3.4). 24 of the edit sites found in coding regions are known to

occur in other taxa. Amino acid altering edit sites were identified in lhbA, petB, ndhD, and accD that were not shared with non-*Marshallia* taxa included in the comparison (Table 3.2).

A number of editing sites were found to exhibit partial editing (Table 3.3). These edits exhibited a wide range (0.22-0.79), encompassing nearly the entire range allowed by the consensus calling parameters. Partial editing was not consistent across the four transcriptomes, with editing percent and the genes being edited varying. Only 2 edit sites were found to be edited to greater than the 80% threshold to preclude partial editing across all four samples. The percent of unedited/partially edited sites within coding regions per sample ranged from 16.29% in *Marhsallia obovata* (M3.9) to as high as 74.19% in *Marshallia mohrii* (M21r) (Table 3.4).

The most common amino acid change was S-L, occurring from 35.48% of edits (Table 5). Leucine was the most common amino acid resulting from edits, occurring from 61.29% of edits (Table 3.6). All but one of the 31 edits occurring in coding regions altered the amino acid sequence by a second position codon alteration, with a single edit occurring in the third position (petB edit 1) (Table 3.7).

Discussion

Of the 31 identified sites within coding regions, 24 were found to be shared with other taxa including *Nicotiana, Arabidopsis, Oryza* and *Pisum sativum*. These sites most likely represent conserved editing sites that evolved early in plant evolution and are present in most Angiosperms. Seven edit sites within coding regions were found to be unique among the model taxa used in the investigation of RNA editing. These sites may represent novel editing sites, but more study within the Asteraceae and other taxa would be required to properly evaluate conservation. No A-I edit was found within the trna-R (ACG) gene, a typically conserved edit.

While this has been found to be a highly conserved editing site, occurring in bacteria to angiosperm plastid genomes, there is some evidence to suggest that it is not necessary nor always found in higher plants (Aldinger *et al* 2012). The absence of this edit would likely be due to a lack of editing at the time of sequencing or fixation of the edit into the genome. To test this, additional DNA and cDNA libraries could be prepared from a given individual for PCR amplification of this locus. PCR amplicons would be sequenced and compared directly to determine whether the edit site is fixed.

During previous assembly and comparison of *Helianthus annuus* and *Lactuca sativa* plastid genomes, several editing sites were identified (Timme *et al* 2007). Edit sites were specifically found in rps3 and psbC that alter the amino acid sequence to create stop codons. These same edits were not identified in this study of the *Marshallia* chloroplast genome. These genes contained the TAA and TGA stop codons, respectively, in both DNA and RNA sequences, suggesting that these sites have become fixed within the *Marshallia* plastid genome. These edits are found in both *Lactuca* and *Helianthus*, with *Lactuca* (tribe Cichorieae) being a more distant relative to *Marshallia* and *Helianthus* (tribe Heliantheae). A number of additional edit sites considered to be conserved that have been identified in the *H. annuus* and *L. sativa* plastid genomes were not identified in *Marshallia*. These include edits in ndhF, psbC, psbZ, rbcL, rpl16, rpl23, rpl32, rps3, and ycf1.

Several editing sites were identified that were not shared with other taxa included in the comparison. These include edits lhbA, petB1, petB2, ndhD5, matK, and accD2. These may represent novel editing sites as they appear to have no taxa sharing them, although a more thorough investigation into plastid RNA editing in more closely related taxa, including *Helianthus* and *Lactuca*, will be required to confirm this.

Putative edit sites were identified in rpoC1 and ycf3 (genes that have been found to be edited in other species), but were found to located near splicing sites, which has been found to cause the appearance of partial editing. Some transcripts were not completely sequenced, so they were not able to be assembled due to low coverage. Other areas were more variable than others and as such were unable to be assessed conservation of edit sites within these regions. To determine the full extent of plastid editing, a systematic comparison of genome and transcriptome of the same individual will be required.

Most editing sites identified exhibited some degree of partial editing in at least one of the four transcriptomes analyzed. Partial editing ranged from ~22% to ~80%, varying from gene to gene and from sample to sample. The percent of unedited/partially edited sites per sample ranged from 16.13% in M3.9 to 74.19% in M21r. Both transcriptome accessions of *Marshallia mohrii* exhibited high levels of partial editing compared to *M. obovata* and *M. trinervia.* While there are sources of noise that could explain the higher levels of partial editing in this analysis, it is unlikely given that these positions were not in close proximity to splice sites or other sources of disagreement. *M. mohrii* is a putative allopolyploid whose parent species are currently undetermined, though it has been thought that one may be *M. trinervia. M. trinervia* and *M. obovata* are both diploid. The great discrepancy in percent of unedited/partially edited sites may have some correlation with *M. mohrii*'s hybridization and polyploidy within the nucleus as the genomic sequences within the chloroplast are highly conserved in this genus. With the edit sites and flanking sequences being shared, the nuclear elements affecting editing may have diverged since the hybridization event. In plants, the nuclear genome and genes typically evolve at a higher rate than the chloroplast genes and genome potentially allowing this type of divergence to occur (Wolfe *et al* 1987). Due to inconsistent nature of partial editing across the 4 samples, it is

likely that the degree to which a given site is edited at any given time is variable, depending on other transcriptional and environmental factors.

Most RNA editing leads to the alteration of an amino acid sequence so that a less hydrophobic residue is replaced by a more hydrophobic residue. The most common edit was found to lead to a change of Serine for Leucine. Leucine is much more hydrophobic that Serine, with hydropathy indices of 3.8 and -0.8, respectively. These edits also produce a more conserved residue when compared to other taxa (Table 2). The most common position to be edited within codons was the 2nd position. 96.77% (30 out of 31) of edits within coding regions occurred in the second position. Only petB edit 1 occurred within the 3rd position, allowing the edit to be silent. These results are very typical, with most known edits within coding regions producing a Leucine by alteration of a 2nd position nucleotide within a codon (Jobson and Qiu 2008).

This work has several important points ranging from demonstrating conservation of editing in non-model taxa to demonstrating the usefulness of NGS technology in identification and quantification of RNA editing. Given that most work has been conducted on a small range of taxa (*Nicotiana*, *Arabidopsis*, pea, rice), this is important for increasing our understanding of the phylogenetic distribution of RNA editing.

Literature Cited

1. Aldinger, C. A., Leisinger, A.-K., Gaston, K. W., Limbach, P. A., & Igloi, G. L. 2012. The absence of A-to-I editing in the anticodon of plant cytoplasmic tRNAArgACG demands a relaxation of the wobble decoding rules. *RNA Biology*, **9**(10), 1239–1246. http://doi.org/10.4161/rna.21839

2. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol*. **215**, 403-410.

3. Bailey, T. L., Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California.

4. Begu, D., Castandet, B. & Araya, A. 2011. RNA editing restores critical domains of a group I intron in fern mitochondria. *Curr Genet* **57**, 317–325.

5. Bock, R., Kossel, H. & Maliga, P. 1994. Introduction of a heterologous editing site into the tobacco plastid genome: the lack of RNA editing leads to a mutant phenotype. *The EMBO Journal* **13,** 4623–4628.

6. Bock, R., & Koop, H.-U. 1997. Extraplastidic site-specific factors mediate RNA editing in chloroplasts. *The EMBO Journal*, **16**(11), 3282–3288.

7. Bock, R. 1998. Analysis of RNA Editing in Plastids. *METHODS: A Companion to Methods in Enzymology* **15,** 75–83.

8. Boisvert, S. Laviolette, F.and Jacques Corbeil. 2010. Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies. *Journal of Computational Biology*. **17**, 1519-1533.

9. Boussardon, C., Salone, V., Avon, A., Berthome, R., Hammani, K., Okuda, K., … Lurin, C. 2012. Two Interacting Proteins Are Necessary for the Editing of the NdhD-1 Site in Arabidopsis Plastids. *The Plant Cell*, **25**, 3684–3694.

10. Chateigner-Boutin, A.-L., & Small, I. 2007. A rapid high-throughput method for the detection and quantification of RNA editing based on high-resolution melting of amplicons. *Nucleic Acids Research*, **35**(17), e114. doi:10.1093/nar/gkm640

11. Chateigner-Boutin, A.-L. & Small, I. 2010. Plant RNA editing. *RNA Biology* **7,** 213–219.

12. Chaudhuri, S., & Maliga, P. 1996. Sequences directing C to U editing of the plastid psbL mRNA are located within a 22 nucleotide segment spanning the editing site. *The EMBO Journal*, **15**(21), 5958–5964.

13. Choury, D. & Araya, A. 2006. RNA editing site recognition in heterologous plant mitochondria. *Curr Genet* **50**, 405–416.

14. Delannoy E, Le Ret M, Faivre-Nitschke E, Estavillo GM, Bergdoll M, Taylor NL, *et al*. 2009. Arabidopsis tRNA adenosine deaminase arginine edits the wobble nucleotide of chloroplast tRNAArg(ACG) and is essential for efficient chloroplast translation. Plant Cell, **21**:2058-71.

15. Dombrovska, O. & Qiu, Y.-L. 2004. Distribution of introns in the mitochondrial gene *nad1* in land plants: phylogenetic and molecular evolutionary implications. *Molecular Phylogenetics and Evolution* **32,** 246–263.

16. Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11-15.

17. Du, P., He, T. & Li, Y. 2007. Prediction of C-to-U RNA editing sites in higher plant mitochondria using only nucleotide sequence features. *Biochemical and Biophysical Research Communications* **358,** 336–341.

18. Du, P. & Li, Y. 2008. Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. *Journal of Theoretical Biology* **253,** 579–586.

19. Duffy, A. M., Kelchner, S. A. & Wolf, P. G. 2009. Conservation of selection on *matK* following an ancient loss of its flanking intron. *Gene* **438,** 17–25.

20. Farre, J.-C., Aknin, C., Araya, A. & Castandet, B. 2012. RNA Editing in Mitochondrial Trans-Introns Is Required for Splicing. *PLOS One* **7**, e52644.

21. Freyer, R., Kiefer-Meyer, M.-C., & Kossel, H. 1997. Occurrence of plastid RNA editing in all major lineages of land plants. *Proc. Natl. Acad. Sci.*, **94**, 6285–6290.

22. Giege, P. & Brennicke, A. 1999. RNA editing in Arabidopsis mitochondria effects 441 C to U changes in ORFs. *PNAS* **96,** 15324–15329.

23. Gouy M., Guindon S. & Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* **27**, 221-224.

24. Hermann, M. & Bock, R. 2000. Transfer of plastid RNA-editing activity to novel sites suggests a critical role for spacing in editing-site recognition. *Proc. Natl. Acad. Sci. USA* **96,** 4856–4861.

25. Hiesel, R., Combettes, B. & Brennicke, A. 1994. Evidence for RNA editing in mitochondria of all major groups of land plants except the Bryophyta. *Proc. Natl. Acad. Sci. USA* **91,** 629–633.

26. Hogg, M., Paro, S., Keegan, L. P. & O'Connell, M. A. 2011. RNA editing by mammalian ADARs. *Adv. Genet.* **73,** 87–120.

27. Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res*., **9**, 868-877.

28. Inada, M., Sasaki, T., Yukawa, M., Tsudzuki, T., & Sugiura, M. 2004. A Systematic Search for RNA Editing Sites in Pea Chloroplasts: an Editing Event Causes Diversification from the Evolutionarily Conserved Amino Acid Sequence. *Plant Cell Physiol.*, **45**(11), 1615–1622.

29. Jin, Y. *et al.* 2007. RNA editing and alternative splicing of the insect nAChR subunit alpha6 transcript: evolutionary conservation, divergence and regulation. *BMC Evolutionary Biology* **7,** 98.

30. Jobson, R. W. & Qiu, Y.-L. 2008. Did RNA editing in plant organellar genomes originate under natural selection or through genetic drift? *Biology Direct* **3**.

31. Karcher, D., & Bock, R. 2009. Identification of the chloroplast adenosine-to-inosine tRNA editing enzyme. *RNA*, **15**, 1251–1257.

32. Katoh, S. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution.* **30**, 772-780.

33. Knoop, V. 2011. When you can't trust the DNA: RNA editing changes transcript sequences. *Cell. Mol. Life Sci.* **68,** 567–586.

34. Kugita, M., Yamamoto, Y., Fujikawa, T., Matsumoto, T. & Yoshinaga, K. 2003. RNA editing in hornwort chloroplasts makes more than half the genes functional. *Nucleic Acids Research* **31,** 2417–2423.

35. Langmead B, Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359.

36. Lavrov, D. V., Brown, W. M. & Boore, J. L. 2000. A novel type of RNA editing occurs in the mitochondrial tRNAs of the centipede Lithobius forficatus. *PNAS* **97,** 13738–13742.

37. Li, M. *et al.* 2011. Widespread RNA and DNA Sequence Differences in the Human Transcriptome. *Science* **333,** 53–58.

38. Liang, G. *et al.* 2013. RNA editing of hepatitis B virus transcripts by activation-induced cytidine deaminase. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 2246–2251.

39. Maier UG, Bozarth A, Funk HT, Zauner S, Rensing SA, Schmitz-Linneweber C, *et al*. 2008. Complex chloroplast RNA metabolism: just debugging the genetic programme? BMC Biol, **6(**36).

40. Malek, O., Lattig, K., Hiesel, R., Brennicke, A. & Knoop, V. 1996. RNA editing in bryophytes and a molecular phylogeny of land plants. *The EMBO Journal* **15,** 1403–1411.

41. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD and Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* **14**, 193-202.

42. Miyata, Y. & Sugita, M. 2004. Tissue- and stage-specific RNA editing of *rps14* transcripts in moss (Physcomitrella patens) chloroplasts. *J. Plant Physiol.* **161,** 113–115.

43. Miyata, Y., Sugiura, C., Kobayashi, Y., Hagiwara, M. & Sugita, M. 2002. Chloroplast ribosomal S14 protein transcript is edited to create a translation initiation codon in the moss Physcomitrella patens. *Biochimica et Biophysica Acta* **1576,** 346–349.

44. Miyamoto, T., Obokata, J. & Gugiura, M. 2002. Recognition of RNA Editing Sites Is Directed by Unique Proteins in Chloroplasts: Biochemical Identification of cis-Acting Elements and trans-Acting Factors Involved in RNA Editing in Tobacco and Pea Chloroplasts. *Mol. Cell. Biol.* **22,** 6726.

45. Mower, J. P. 2005. PREP-Mt: predictive RNA editor for plant mitochondrial genes. *BMC Bioinformatics* **9.**

46. Mower, J. P. 2009. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acid Research* **37**.

47. Mulligan, R. M., Williams, M. A. & Shanahan, M. T. 1999. RNA Editing Site Recognition in Higher Plant Mitochondria. *The American Genetic Association* **90,** 338–344.

48. Mulligan, R. M., Williams, M. A. & Shanahan, M. T. 1999. RNA Editing Site Recognition in Higher Plant Mitochondria. *The American Genetic Association* **90,** 338–344.

49. Neuwirt, J., Takenaka, M., Vand Der Merwe, J. A. & Brennicke, A. 2005. An in vitro RNA editing system from cauliflower mitochondria: Editing site recognition parameters can vary in different plant species. *RNA* **11,** 1563–1570.

50. Nugent, J. M. & Palmer, J. D. 1991. RNA-Mediated Transfer of the Gene coxII from the Mitochondrion to the Nucleus during Flowering Plant Evolution. *Cell* **66,** 473–481.

51. Okuda, K., Chateigner-Boutin, A.-L., Nakamura, T., Delannoy, E., Sugita, M., Myouga, F., … Shikanai, T. 2009. Pentatricopeptide Repeat Proteins with the DYW Motif Have Distinct Molecular Functions in RNA Editing and RNA Cleavage in Arabidopsis Chloroplasts. *The Plant Cell*, **21**, 146–156.

52. Okuda, K., Hammani, K., Tanz, S. K., Peng, L., Fukao, Y., Myouga, F., … Shikanai, T. 2010. The pentatricopeptide repeat protein OTP82 is required for RNA editing of plastid ndhB and ndhG transcripts. *The Plant Journal*, **61**, 339–349. doi:10.1111/j.1365-313X.2009.04059.x

53. Ruwe, H., Kupch, C., Teubner, M. & Schmitz-Lenneweber, C. 2011. The RNA-recognition motif in chloroplasts. *Journal of Plant Physiology* **168,** 1361–1371.

54. Salone, V., Rudinger, M., Polsakiewicz, M., Hoffmann, B., Groth-Malonek, M., Szurek, B., … Lurin, C. 2007. A hypothesis on the identification of the editing enzyme in plant organelles. *FEBS Letters*, **581**, 4132–4138.

55. Schuster, W., Hiesel, R., Wissinger, B. and Brennicke, A. 1990. RNA editing in the cytochrome b locus of the higher plant Oenothera berteriana includes a U-to-C transition. *Molecular and Cellular Biology* **10,** 2428–2431.

56. Sper-Whitis, G. L., Moody, J. L. & Vaugn, J. C. 1996. Universality of mitochondrial RNA editing in cytochrome-c oxidase subunit I (*coxI*) among the land plants. *Biochimica et Biophysica Acta* **1307,** 301–308.

57. Takenaka, M., Verbitskiy, D., van der Merwe, J. A., Zehrmann, A. & Brennicke, A. 2008. The process of RNA editing in plant mitochondria. *Mitochondrion* **8,** 35–46.

58. Takenaka, M., Zehrmann, A., Verbitskiy, D., Hartel, B. & Brennicke, A. 2013. RNA Editing in Plants and Its Evolution. *Annu. Rev. Genet.* **47,** 335–352.

59. Tangphatsornruang, S. *et al.* 2011. Characterization of the complete chloroplast genome of Hevea brasiliensis reveals genome rearrangement, RNA editing sites and phylogenetic relationships. *Gene* **475,** 104–112.

60. Tasaki, E., & Sugita, M. 2010. The moss Physcomitrella patens, a model plant for the study of RNA editing in plant organelles. *Plant Signaling & Behavior*, **5**(6), 727–729.

61. Timme, R. E., Kuehl, J. V., Boore, J. L., & Jansen, R. K. 2007. A COMPARATIVE ANALYSIS OF THE LACTUCA AND HELIANTHUS (ASTERACEAE) PLASTID GENOMES: IDENTIFICATION OF DIVERGENT REGIONS AND CATEGORIZATION OF SHARED REPEATS. *American Journal of Botany*, **94**(3), 302–317.

62. Tsudzuki, T., Wakasugi, T., & Sugiura, M. 2001. Comparative Analysis of RNA Editing Sites in Higher Plant Chloroplasts. *J Mol Evol*, **53**, 327–332.

63. Vapnik, V. N., & Vapnik, V. 1998. *Statistical learning theory* (Vol. 2). New York: Wiley.

64. Volchkova, V. A., Dolnik, O., Martinez, M. J., Reynard, O. & Volchkov, V. E. 2011. Genomic RNA editing and its impact on Ebola virus adaptation during serial passages in cell culture and infection of guinea pigs. *J. Infect. Dis.* **204 Suppl 3,** S941–946.

65. Wolf, P. G., Rowe, C. A. & Hasebe, M. 2004. High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern Adiantum capillus-veneris. *Gene* **339,** 89–97.

66. Wolf, P. G. *et al.* 2005. The first complete chloroplast genome sequence of a lycophyte, Huperzia lucidula (Lycopodiaceae). *Gene* **350,** 117–128.

67. Wolfe, K. H., Li, W.-H., & Sharp, P. M. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. USA*, **84**, 9054–9058.

68. Yagi, Y. *et al.* 2013. Pentatricopeptide repeat proteins involved in plant organellar RNA editing. *RNA Biol* **10,** 1419–1425.

69. Yu, Q.-B., Jiang, Y., Chong, K., & Yang, Z.-N. 2009. AtECB2, a pentatricopeptide repeat protein, is required for chloroplast transcript accD RNA editing and early chloroplast biogenesis in Arabidopsis thaliana. *The Plant Journal*, **59**, 1011–1023.

70. Zhang, J., Ruhlman, T. A., Mower, J. P. & Jansen, R. K. 2013. Comparative analyses of two Geraniaceae transcriptomes using next-generation sequencing. *BMC Plant Biology* **13**.

71. Zhu, Y. *et al.* 2014. Abundant and selective RNA-editing events in the medicinal mushroom Ganoderma lucidum. *Genetics* **196,** 1047–1057.

Figures and Tables

Table 3.1. Editing sites identified within coding regions of the *Marshallia* plastid genome.

Gene	Nucleotide Position	Codon	Residue Position	Residue Change
accD	779	tCg	260	S-L
accD	1388	aCt	463	T-I
lhbA	50	tCa	17	S-L
matK	704	tCt	235	S-F
ndhA	107	cCt	36	P-L
ndhA	1073	tCc	358	S-F
ndhB	149	tCa	50	S-L
ndhB	467	cCa	156	P-L
ndhB	586	Cat	196	H-Y
ndhB	737	cCa	246	P-L
ndhB	746	tCt	249	S-F
ndhB	830	tCa	277	S-L
ndhB	836	tCa	279	S-L
ndhB	1481	cCa	494	P-L
ndhD	2	aCg	1	T-M
ndhD	383	tCa	128	S-L
ndhD	599	tCa	200	S-L
ndhD	878	tCa	293	S-L
ndhD	887	cCc	296	P-L
ndhD	1310	tCa	435	S-L
ndhG	347	cCa	116	P-L
petB	12	gtC	4	V-V
petB	418	Cgg	140	R-W
petB	611	cCa	204	P-L
petL	5	cCt	2	P-L
psaI	85	tCa	29	H-Y
psbF	77	tCt	26	S-F
psbL	2	aCg	1	T-M
rps2	134	aCa	45	T-I
rps2	248	tCa	83	S-L
rps14	80	tCa	27	S-L

Table 3.2. Comparison of editing sites identified in *Marshallia* spp. compared to sites experimentally identified in other taxa. X represents an edit shared with *Marshallia* spp.

Gene	*Arabidopsis*	*Nicotiana*	*Pisum sativum*	*Oryza*	*Zea mays*
accD			X		
accD					
lhbA					
matK					
ndhA					
ndhA		X		X	X
ndhB	X	X	X		
ndhB	X	X		X	X
ndhB	X	X	X	X	X
ndhB		X	X	X	X
ndhB	X	X	X		
ndhB	X	X	X	X	X
ndhB	X	X	X	X	
ndhB	X	X		X	X
ndhD	X	X	X		
ndhD	X	X	X		
ndhD		X			
ndhD	X			X	X
ndhD	X				
ndhD					
ndhG				X	
petB					
petB					
petB		X	X		X
petL			X		
psaI		X			
psbF	X		X		
psbL		X			
rps2		X	X	X	
rps2		X	X		
rps14	X		X	X	X

Table 3.3. Base present in alignment per site per sample for coding regions. Proportion of reads with edited base is listed for any position for which an ambiguity symbol was called. - denote positions for which there was no data, * denotes position for which a non-ambiguous edited base was called, C denoted a position for which an unambiguous unedited C was called.

	M2.9	M3.9	M20r	M21r
accD	-	*	*	*
accD	-	*	0.22	*
lhbA	0.4	*	0.72	0.66
matK	-	*	0.45	0.67
ndhA	0.6	*	0.78	*
ndhA	0.79	*	0.71	0.77
ndhB	*	*	0.64	0.64
ndhB	*	*	0.64	0.6
ndhB	*	*	0.57	0.65
ndhB	*	*	0.72	*
ndhB	*	*	0.71	*
ndhB	*	*	0.77	0.25
ndhB	*	*	0.78	0.67
ndhB	-	*	*	0.75
ndhD	*	C	0.21	0.43
ndhD	*	0.75	0.67	0.54
ndhD	*	0.75	0.65	C
ndhD	*	*	0.75	*
ndhD	*	*	0.73	0.79
ndhD	0.5	*	0.68	0.5
ndhG	*	*	0.71	*
petB	0.5	0.63	0.32	C
petB	*	*	*	0.79
petB	*	*	*	0.73
petL	0.75	*	0.62	0.75
psaI	*	0.77	0.71	0.56
psbF	*	0.74	*	0.78
psbL	*	*	*	*
rps14	0.5	*	0.75	*
rps2	-	*	*	0.5
rps2	-	*	*	0.67

Table 3.4. Percent unedited and partially edited sites per sample.

Sample	% sites
M2.9	28
M3.9	16.13
M20.2	70.97
M20.1	74.19

Table 3.5. Proportion of amino acid change types resulting from editing.

Change	N	Proportion
S-L	11	0.354839
S-F	4	0.129032
H-Y	2	0.064516
P-L	8	0.258065
R-W	1	0.032258
T-M	2	0.064516
T-I	2	0.064516
V-V	1	0.032258

Table 3.6. Proportion of amino acids resulting from edits.

Amino Acid	N	Proportion
L	19	0.612903
F	4	0.129032
M	2	0.064516
I	2	0.064516
W	1	0.032258
Y	2	0.064516
silent	1	0.032258

Table 3.7. Proportion of codon positions featuring an editing site.

Position	N	Proportion
1	0	0
2	30	0.967742
3	1	0.032258

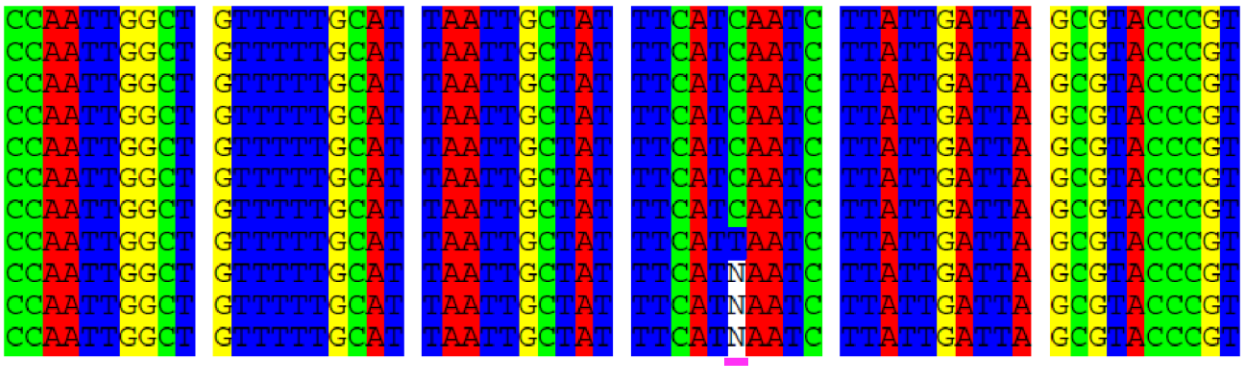

Figure 3.1. Example of genomic and transcriptomic alignments located within the lhbA gene. Sequences 1-7 represent genomic DNA and 8-11 represent RNA. This putative edit site, underlined in pink, shows C's in position 36 in DNA sequences and a T and N's in the RNA.

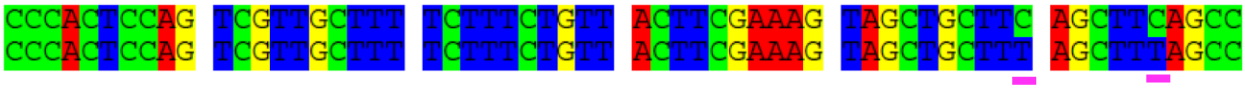

Figure 3.2. Nucleotide alignment of *M. obovata* ndhB genomic sequence and *M. mohrii* (M21r) ndhB transcript showing positions 781 to 840, including two edits sites, ndhB 6 and 7. These sites are underlined in pink.

Figure 3.3. ndhD protein alignment of sequences inferred from *M. obovata* DNA and *M. mohrii* RNA sequences showing ndhB edit sites 4, 5, 6, and 7. These sites are underlined in pink.

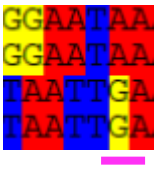

Figure 3.4. Genomic and transcriptomic stop codons for rps3 (top) and psbC (bottom) in M3 (1st and 3rd sequences) and M21r (2nd and 4th sequences). Stop codons are underlined in pink.

Chapter 4. Overview

This work has been conducted under a theme of bioinformatics in a non-model system. A Research has been conducted on members of the genus *Marshallia* (Shreb.) that ranges from investigations into a single gene to genome wide analyses. This work utilized a variety of techniques from expressions analyses and qPCR to bioinformatics techniques and genome assembly software.

Knowledge of Cytokinin Response Factor (CRF) genes have been furthered through investigations using a non-model organism. Novel Clade III CRF motifs were identified through alignment and analysis of 50 Asterid amino acid sequences. These motifs were highly supported by MEME analysis and may represent conserved functional domain. Full transcripts were assembled from two species of *Marshallia* (*M. caespitosa* and *M. mohrii*). These were found to be orthologues of *Sl*CRF5 and showed strong conservation of functional domains. *Mc*CRF3 was also found to be up-regulated by similar stimuli as the Clade III CRFs of *Arabidopsis* and *S. lycopersicum*.

Many conserved and several novel RNA edit sites were identified within the *Marshallia* plastid genome. Most notable were the lack of support for specific edits that are thought to be highly conserved. No evidence of an A-to-I edit in the tRNA-ARG was found. It was also found that edits typically required to create stop codons in rps3 and psbC had been "fixed" in the *Marshallia* chloroplast genomes. Other results of interest include higher proportions of sites being partially edited in *M. mohrii*, a putative allopolyploid.

My thesis work has accomplished several important goals: 1) it has given me an extensive toolbox from which I can build upon in my PhD work, 2) it has expanded our

knowledge of Clade III CRFs outside of model systems, and 3) has extended out knowledge of the chloroplast genome of *Marshallia* and of RNA editing in closely related taxa.