An Investigation of Species Relationships within *Marshallia* Schreb. [Asteraceae] and Mitochondrial Genomes within Poaceae

by

Nathan Daniel Hall

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 5, 2018

Keywords: *Marshallia*, Asteraceae, plastid, *Eleusine,* Poaceae, mitochondria

Approved by

Leslie R. Goertzen, Director of the John D. Freeman Herbarium in the Auburn University Museum of Natural History, Associate Professor of Biology
Robert S. Boyd, Undergraduate Program Officer in the Department of Biological, Professor of Biology
Aaron Rashotte, Associate Professor of Biology
Scott R. Santos, Professor of Biology

Abstract

Genomic sequencing has become ubiquitous within the discipline of Biology. Here I use two case studies in which whole genomic DNA facilitates answers to foundational biological questions for organisms that do not have assembled genomes and have benefited little, if at all, from any sequencing in the past. *Eleusine indica* (L.) Gaertn. (Poaceae) is a prolific weed that is also a known genome donor for *E. coracana* (L.) Gaertn., a subsistence crop used in west Africa and east India. The genus *Marshallia* Schreb. (Asteraceae) comprises eight species, occurs throughout the southeast and is notable for inflorescences with a pink or white puffball appearance. In Chapter 1 general context and background is provided regarding both *E. indica* and *Marshallia* and the aims of the study are provided. Chapter 2 is an assembly of the mitochondrial genome of *Eleusine indica* and a survey of the mitochondrial gene content of Poaceae. Plant mitochondrial genomes are challenging to assemble because of large recombinogenic regions, this in turn makes surveys of mitochondrial gene content hard to accomplish without the use of extensive assembly efforts or laborious wet lab techniques. A new *in silico* approach is described and applied to determine mitochondrial gene content within publicly available datasets. The results are confirmed by past studies, but also expand sampling within Poaceae and provide new insight into the specific timing of mitochondrial gene loss in Poaceae. In Chapter 3 a repetitive element is assembled and characterized as a proof of concept for the usefulness of whole genomic low coverage sequence data within *Marshallia*. In Chapter 4

several facets of these data are employed to determine interspecies relationships within *Marshallia.* Seventeen datasets are used in phylogenetic analysis of assemble plastid genomes, entire ribosomal cistrons and mitochondrial sequences. Additionally the abundance of the repetitive element *Marobo* is characterized in Chapter 3 along with the relative abundance of several repetitive element superfamilies. Finally, a reduced set of transcriptomes is assembled and used for phylogenetic analysis. The results of all assemblies are compared and this yields the first in-depth exploration of the phylogenetic history of the species within *Marshallia*. Relationships within the *Marshallia* show species-specific divergence of the ribosomal DNA, which are in contrast to the interspecific sharing of cytotypes, this pattern is indicative of past introgression events.

Acknowledgments

The completion of this dissertation would have been impossible without the support of many. I could not have achieved this without the guidance and ever evolving curiosity of my advisor Les Goertzen. Your willingness to chat and insistence on starting at first principles have given me clarity of thought and helped organize my biological thinking. Additionally, I am indebted to my committee, for their work with me and willingness provide valuable insights into the worlds of plants and genomic data. With regard to Chapter 2, I gratefully acknowledge the assistance, keen insights, and collaboration of Jeffery Mower University of Nebraska-Lincoln particularly in regard to mitochondrial circularization and the validation of our heat map approach. I also gratefully acknowledge the encouraging and insightful comments of Han Ong, King University. With regard to Chapter 3 I appreciate the careful reading and feedback given by Kate Hertweck of University of Texas at Tyler and Zdenek Kuab, Laboratory of Plant Developmental Genetics, Institute of Biophysics ASCR, Czech Republic. Additionally, I am indebted to my lab mates Curtis Hansen, Anthony Melton, and Morgan Dalis.

I am deeply indebted to my family. My wife who has supported me through this process, and listened to more talk about programming and Linux than is strictly healthy. My In-Laws who have helped with child care, provided meals, learned more biology than they expected and generally supported my family and me. My children who are patiently waiting for us to get a puppy.

Table of Contents

List of Tables

Chapter 2 The mitochondrial genome of *Eleusine indica* and improved characterization of gene loss within Poaceae

Chapter 3 Sequencing and characterization of the *Del/Tekay* Chromovirus family in *Marshallia obovata* (Asteraceae)

Chapter 4 An investigation of interspecies relationships within *Marshallia*

List of Illustrations

Chapter 2 The mitochondrial genome of *Eleusine indica* and improved characterization of gene loss within Poaceae

Chapter 3 Sequencing and characterization of the *Del/Tekay* Chromovirus family in *Marshallia obovata* (Asteraceae)

Chapter 4 An investigation of interspecies relationships within *Marshallia*

List of Abbreviations

BEP	Bambusoideae, Ehrhartoideae and Pooideae
DNA	deoxyribonucleic acid
dN	non-synonymous substitution rate
dS	synonymous substitution rate
DUF	domain of unknown function
GATK	Genome Analysis Toolkit
GPWG	Grass Phylogeny Working Group
HSD	honest significant difference
ITS	internal transcribed spacer
LINE	long interspersed nuclear elements
LTR	long terminal repeat
mt	mitochondrial
NCBI	the National Center for Biotechnology Information
PACMAD	Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae and Danthonioideae
PCA	principal component analysis
ORF	open reading frame
rDNA	ribosomal DNA
RNA	ribonucleic acid

SNP	single nucleotide polymorphism
tRNA	transfer RNA

Chapter 1 Introduction to study systems

General Introduction

The discipline of Biology has experienced an explosion of available data and resources with which to work within the last decade (Eisenberg et al. 2000; O'Driscoll et al. 2013). These resources have aided research efforts on traditionally non-model organisms, i.e., organisms without sequenced genomes or robust resources created by decades of research (Ellegren 2014; Cammen et al. 2016; Therkildsen and Palumbi 2017). The judicious use of sequencing and assembly methods has allowed researchers to resolve or clarify previously intractable questions by bringing large quantities of hiqh quality data to bear. As examples, Wickett et al. (2014) used the largest assembly of nuclear genes, at the time of publishing, to test hypotheses about the origins of land plants. The Grass Phylogeny Working (2012) group addressed the finer scale phylogenetic question of C4 in photosynthesis in grasses. Finally, several researchers have addressed phylogenetic (Steele et al. 2012; Besnard et al. 2013; Malé et al. 2014; Hardion et al. 2017) and physiological (Shi et al. 2014; Ranjan et al. 2014; Luria et al. 2014; Bushman et al. 2016; Skorupa et al. 2016) questions.

Sequencing is increasing and the data produced often hold the promise of utility beyond the intended research aims that produced them. It is important to test how well and to what degree sequencing and assembly approaches resolve hard problems in biology and to what degree data produced by other research endeavors can be used to advance science (Eisenberg et al. 2000; Philip Chen and Zhang 2014). Three hard problems to consider are: 1) resolving

inter-species relationships within a genus containing closely-related species, 2) assembly of a plant mitochondrial genome, and 3) characterization of plant mitochondrial gene content.

Low coverage sequencing approaches are often used to resolve phylogenetic relationships at several levels of taxonomic interest (Kane et al. 2012; McKain et al. 2012; Besnard et al. 2013; Bock et al. 2014; Panero et al. 2014). Testing the resolving power of low-coverage sequencing approaches uses two sequencing approaches, genome skimming (Straub et al. 2012) and transcriptome assembly (Grabherr et al. 2011). The aim in this case study is to compare the resolving power of these approaches on closely related species in a previously intractable genus, *Marshallia* (Watson and Estes 1990; Watson et al. 1994).

The high coverage illumina sequencing found in genome projects which ranges of *ca*. 10x to 150x (Ekblom and Wolf 2014) are useful for characterizing plant mitochondrial genomes in addition to nuclear genomes (Clifton et al. 2004; Cui et al. 2009). To test the efficacy of these sequencing methods within the context of *Eleusine indica,* a mitochondrial genome will be assembled.

Finally, the data extracted for high coverage and low coverage genome sequencing often have utility that transcends the initial aims of the research such as determining mitochondrial gene content (Straub et al. 2012) and determining repetitive element content (Novák et al. 2010; Staton and Burke 2015). Southern blotting, slot blots and targeted polymerase chain reaction paired with sequencing have traditionally been used to determine mitochondrial gene content when whole mitochondrial genomes have not been available (Palmer et al. 2000; Adams et al. 2001; Bergthorsson et al. 2003; Ong and Palmer 2006; Wu et al. 2017). My first aim for these datasets is to produce an *in silico* southern blot to determine mitochondrial gene content within Poaceae, a taxon that has been studied extensively, but not exhaustively and recent work has

focused mainly on one gene or cistron at a time (Atluri et al. 2015; Wu et al. 2017). An *in silico* approach would allow for multiple tests to be run simultaneously and for the rapid application of techniques to new data as they become available. The final aim, to determine the repetitive element fraction, will be carried out with *Marshallia* in two steps. First, a single repetitive element will be sequenced and annotated, to provide a basic view of the completeness of the data. Second, a read clustering algorithm (Staton and Burke 2015) will be employed to determine the percent of each genome that is composed of moderate to highly repetitive sequences.

Organisms

Eleusine indica

Eleusine indica (L.) Gaertn. is a cosmopolitan weed of economic importance. In addition to being widespread, it commonly occurs with *Eleusine coracana,* a valuable food crop that was domesticated in western Africa and has been a staple crop there and in eastern India for 5,000 and 3,000 years, respectively (Neves et al. 2005). *Eleusine indica* is known to be a progenitor of the allotetraploid *E. corocana* (Hittalmani et al. 2017). The genus *Eleusine* comprises 8-10 species and is generally restricted to Africa, with notable exceptions of *E. indica*, *E. coracana*, and *E. tristachya,* which occurs in South America (Phillips 1972; Hilu and Johnson 1997; Neves et al. 2005). The genus contains both annuals and perennials, yet this appears to be a polyphyletic trait (Hilu and Johnson 1997). *Eleusine* is in tribe Choridoidea, subfamily PACMAD of Poaceae (Peterson et al. 2001; Grass Phylogeny Working Group II 2012).

Eleusine indica is between 15 and 85 cm tall and produces spikes that are 3-7mm in length with blackish elliptical to oblong grain that is never exposed at maturity (Phillips 1972). This contrasts with *E. coracana* which is also annual but its spikes are approximately two times larger and grain is brown, globose, and often exposed at maturity (Phillips 1972)**.**

Marshallia

Marshallia Schreb. [Asteraceae:Helenieae] is a southeastern endemic genus of woodland wildflowers that provides an excellent test case for the utility of low coverage high throughput sequencing, phylogenomic techniques, and specimen informatics on non-model systems. The genus *Marshallia* has been notoriously intractable to past phylogenetic investigations (Smith and Shine 1998) owing to inter-gradations of morphological characters and uniformity of conserved molecular markers (Watson and Estes 1990). These factors have confounded attempts to produce a reliable phylogeny of *Marshallia*. Work by Channell (1957) notes the lack of defining characters among species. Recent work by Hansen and Goertzen (2014) has shown that lineages can be separated via ITS barcoding but phylogenetic relationships remain unresolved.

Marshallia contains 8 well-defined species (*sensu* Hansen and Goertzen 2015): *M. mohrii* Beadle & F.E. Boynton, *M. trinervia* (Walter) Trel., *M. grandiflora* Beadle & F.E. Boynton, *M. graminifolia* (Walter) Small, *M. caespitosa* Nutt. ex DC., *M. legrandii* Weakley, *M. ramosa* Beadle & F.E. Boynton, and *M. obovata* (Walter) Beadle & F.E. Boyton. *Marshallia* ranges from western Pennsylvania to east Oklahoma and Texas. They are found throughout the southeastern coastal plains and into the Appalachian Mountains (Channell 1957). Plants from the genus *Marshallia* were first described in 1788 by Schreber who assigned it to *Athanasia* (Channell 1957). The group was later given its own genus name because of its distinct floral characteristics. *Marshallia* was initially a degenerate genus name, and assignment of *Marshallia* to the New World plants would not be made final until 1842 (Channell 1957). *Marshallia* are scapose, sub-scapose, and caulescent in growth form, with single to many discoid inflorescences containing flowers that have unusually elongated floral tubes (Channell 1957; Weakley and Poindexter 2012). The unique floral structure results in a white (or pink) puffball-like appearance. Even

though there are many graded characters among congenerics, each species possesses a suite of characters that have been used to circumscribe it. *Marshallia mohrii* grows on limestone, sandstone, dolomite glades, and calcareous prairies (Weakley and Poindexter 2012); it is also known to branch more frequently than other species (Channell 1957; Weakley and Poindexter 2012). It has gained additional attention for its threatened status as a Federally-listed woodland herb (U.S. Fish and Wildlife 1988). *Marshallia trinervia,* possesses three-nerved or veined leaves and grows in rocky shallow calcareous moist soil (Weakley and Poindexter 2012). *Marshallia grandiflora* bears a striking resemblance to *M. trinervia.* However, *M. grandiflora* exhibits a decrease in leaf size inverse to leaf height, that is, the higher the leaf on the stem the smaller it is, and it also exhibits branching more frequently than *M. trineriva* (Channell 1957). In contrast, plants of *M. legrandii* possess a solitary head, are the tallest in the genus at 6-9 dm tall (Weakley and Poindexter 2012), and have a greatly restricted distribution. *Marshallia ramosa* has one to multiple heads but occurs on ultramafic outcrops, pinelands, and grit glades. In this treatment of *Marshallia* I have combined *M. tenuifolia* (Walter) Small subsp. *tenuifolia* (Raf.) L. Watson with *M. graminifolia*. The separation of these species is likely a result of geography and not convincing morphology or genetic differences (Watson and Estes 1990; Watson et al. 1994; Hansen and Goertzen 2014). *Marshallia caespitosa* is at the western edge of *Marshallia* distribution and often produces many inflorescences (Channell 1957). The species also contains known variations in ploidy, likely representing an autopolyploid variety (Watson and Estes 1990). *Marshallia obovata* has an Appalachian Mountains and sandy piedmont distribution and possesses obovate palea (Channell 1957).

Past systematic work on the genus has largely been carried out by phenetic analysis (Watson and Estes 1990), chromosome counts (Watson and Estes 1987; Watson et al. 1991a),

classic genetic markers (Watson et al. 1994), and ITS markers (Hansen and Goertzen 2014). Tribal placement of this genus has been carried out using restriction sites (Watson et al. 1991b) and integrated approaches (Panero et al. 2014). The most recent tribal placement (Panero et al. 2014) of *Marshallia* puts it with Asteroidea [Asteraceae]. Baldwin and Wessa (2000) placed *Marshallia* as a basal taxon in subtribe Helenieae. Their placement was supported by sequence alignments of 18S-26S rDNA ITS region sequences. Broader analysis of Heliantheae reveals that Helenieae is basal within Heliantheae (Goertzen et al. 2003). All these analyses taken together provide a picture of *Marshallia* as basal to all Heliantheae and very isolated from other taxa.

Past approaches to resolve the phylogeny of congenerics within *Marshallia* have been thwarted by lack of phylogenetic signal within selected features (Watson and Estes 1990; Watson et al. 1991a, b, 1994).

Aims and Where to Find Them.

Chapter 2 contains the assembly of the *Eleusine indica* mitochondrial genome and test of mitochondrial gene content. Chapter 3 contains the fine scale assembly and annotation of a repetitive element within *Marshallia*. Chapter 4 contains the assembly and analysis of low coverage high-throughput sequencing and the use of read mapping and graph based clustering to sample the repetitive fraction of the genomes within *Marshallia*.

References

Adams KL, Rosenblueth M, Qiu YL, Palmer JD (2001) Multiple losses and transfers to the nucleus of two mitochondrial succinate dehydrogenase genes during angiosperm evolution. Genetics 158:1289–1300

Atluri S, Rampersad SN, Bonen L (2015) Retention of functional genes for S19 ribosomal protein in both the mitochondrion and nucleus for over 60 million years. Mol Genet Genomics 290:2325–2333

Baldwin BG, Wessa BL (2000) Phylogenetic placement of *Pelucha* and new subtribes in Helenieae sensu stricto (Compositae). Syst Bot 25:522–538

Bergthorsson U, Adams KL, Thomason B, Palmer JD (2003) Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature 424:197–201

Besnard G, Christin P-A, Malé P-JG, et al. (2013) Phylogenomics and taxonomy of *Lecomtelleae* (Poaceae), an isolated panicoid lineage from Madagascar. Ann Bot 112:1057–1066

Bock DG, Kane NC, Ebert DP, Rieseberg LH (2014) Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. New Phytol 201:1021–1030

Bushman BS, Amundsen KL, Warnke SE, et al. (2016) Transcriptome profiling of Kentucky bluegrass (*Poa pratensis* L.) accessions in response to salt stress. BMC Genomics 17:48

Cammen KM, Andrews KR, Carroll EL, et al. (2016) Genomic methods take the plunge: recent advances in high-throughput sequencing of marine mammals. J Hered 107:481–495

Channell RB (1957) A revisional study of the genus *Marshallia* (Compositae). Contributions from the Gray Herbarium of Harvard University 41–130

Clifton SW, Minx P, Fauron CM-R, et al. (2004) Sequence and comparative analysis of the maize NB mitochondrial genome. Plant Physiol 136:3486–3503

Cui P, Liu H, Lin Q, et al. (2009) A complete mitochondrial genome of wheat (*Triticum aestivum* cv. Chinese Yumai), and fast evolving mitochondrial genes in higher plants. J Genet 88:299–307

Eisenberg D, Marcotte EM, Xenarios I, Yeates TO (2000) Protein function in the post-genomic era. Nature 405:823–826

Ekblom R, Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026–1042

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol 29:51–63

Goertzen LR, Cannone JJ, Gutell RR, Jansen RK (2003) ITS secondary structure derived from comparative analysis: implications for sequence alignment and phylogeny of the Asteraceae. Mol Phylogenet Evol 29:216–234

Grabherr MG, Haas BJ, Yassour M, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

Grass Phylogeny Working Group II (2012) New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. New Phytol 193:304–312

Hansen CJ, Goertzen LR (2014) Validation of nrDNA ITS as a DNA barcode for *Marshallia* (Asteraceae). Paysonia 3:5–10

Hardion L, Verlaque R, Haan-Archipoff G, et al. (2017) Cleaning up the grasses dustbin: systematics of the Arundinoideae subfamily (Poaceae). Plant Syst Evol 303:1331–1339

Hilu KW, Johnson JL (1997) Systematics of *Eleusine* Gaertn. (Poaceae: Chloridoideae): chloroplast DNA and total evidence. Ann Mo Bot Gard 84:841–847

Hittalmani S, Mahesh HB, Shirke MD, et al. (2017) Genome and transcriptome sequence of Finger millet (*Eleusine coracana* (L.) Gaertn.) provides insights into drought tolerance and nutraceutical properties. BMC Genomics 18:465

Kane N, Sveinsson S, Dempewolf H, et al. (2012) Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. Am J Bot 99:320–329

Luria N, Sela N, Yaari M, et al. (2014) De-novo assembly of mango fruit peel transcriptome reveals mechanisms of mango response to hot water treatment. BMC Genomics 15:957

Malé P-JG, Bardon L, Besnard G, et al. (2014) Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. Mol Ecol Resour 14:966–975

McKain MR, Wickett N, Zhang Y, et al. (2012) Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). Am J Bot 99:397–406

Neves SS, Swire-Clark G, Hilu KW, Baird WV (2005) Phylogeny of *Eleusine* (Poaceae: Chloridoideae) based on nuclear ITS and plastid trnT-trnF sequences. Mol Phylogenet Evol 35:395–419

Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:378

O'Driscoll A, Daugelaite J, Sleator RD (2013) "Big data", Hadoop and cloud computing in genomics. J Biomed Inform 46:774–781

Ong HC, Palmer JD (2006) Pervasive survival of expressed mitochondrial *rps14* pseudogenes in grasses and their relatives for 80 million years following three functional transfers to the nucleus. BMC Evol Biol 6:55

Palmer JD, Adams KL, Cho Y, et al. (2000) Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc Natl Acad Sci U S A 97:6960–6966

Panero JL, Freire SE, Ariza Espinar L, et al. (2014) Resolution of deep nodes yields an improved backbone phylogeny and a new basal lineage to study early evolution of Asteraceae. Mol Phylogenet Evol 80:43–53

Peterson PM, Soreng RJ, Davidse G, et al. (2001) Catalogue of New World Grasses (Poaceae): II. subfamily Chloridoideae. Contrib U S Natl Herb 41:1–255

Philip Chen CL, Zhang C-Y (2014) Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. Inf Sci 275:314–347

Phillips SM (1972) A Survey of the Genus *Eleusine* Gaertn. (Gramineae) in Africa. Kew Bull 27:251–270

Ranjan A, Ichihashi Y, Farhi M, et al. (2014) De novo assembly and characterization of the transcriptome of the parasitic weed dodder identifies genes associated with plant parasitism. Plant Physiol 166:1186–1199

Shi H, Jiang C, Ye T, et al. (2014) Comparative physiological, metabolomic, and transcriptomic analyses reveal mechanisms of improved abiotic stress resistance in bermudagrass [*Cynodon dactylon* (L). Pers.] by exogenous melatonin. J Exp Bot 66:681–694

Skorupa M, Gołębiewski M, Domagalski K, et al. (2016) Transcriptomic profiling of the salt stress response in excised leaves of the halophyte *Beta vulgaris* ssp. *maritima*. Plant Sci 243:56–70

Smith S, Shine C (1998) Plate 343. *Marshallia grandiflora*. Curtis's Botanical Magazine 15:158-163

Staton SE, Burke JM (2015) Transposome: a toolkit for annotation of transposable element families from unassembled sequence reads. Bioinformatics 31:1827–1829

Steele PR, Hertweck KL, Mayfield D, et al. (2012) Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. Am J Bot 99:330–348

Straub SCK, Parks M, Weitemier K, et al (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am J Bot 99:349–364

Therkildsen NO, Palumbi SR (2017) Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Mol Ecol Resour 17:194–208

U.S. Fish and Wildlife (1988) Endangered and threatened wildlife and plants; determination of *Marshallia mohrii* (Mohr's Barbara's buttons) to be a threatened species. Fed Regist 53:4698–34701

Watson LE, Elisens WJ, Estes JR (1994) Genetic variation within and among populations of the *Marshallia graminifoli*a complex (Asteraceae). Biochem Syst Ecol 22:577–582

Watson LE, Elisens WJ, Estes JR (1991a) Electrophoretic and cytogenetic evidence for allopolyploid origin of *Marshallia mohrii* (Asteraceae). Am J Bot 78:408–416

Watson LE, Estes JR (1987) Chromosomal evolution of *Marshallia* (Asteraceae). Am J Bot 74:764

Watson LE, Estes JR (1990) Biosystematic and phenetic analysis of *Marshallia* (Asteraceae). Syst Bot 15:403–414

Watson LE, Jansen RK, Estes JR (1991b) Tribal placement of *Marshallia* (Asteraceae) using chloroplast dna restriction site mapping. Am J Bot 78:1028–1035

Weakley AS, Poindexter DB (2012) A new species of *Marshallia* (Asteraceae, Helenieae, Marshalliinae) from mafic woodlands and barrens of North Carolina and Virginia. Phytoneuron 105:1-17

Wickett NJ, Mirarab S, Nguyen N, et al. (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci USA 111:E4859–68

Wu Z, Sloan DB, Brown CW, et al. (2017) Mitochondrial retroprocessing promoted functional transfers of *rpl5* to the nucleus in grasses. Mol Biol Evol 34:2340–2354

Chapter 2 The mitochondrial genome of *Eleusine indica* and improved characterization of gene loss within Poaceae

Abstract

Plant mitochondrial genomes are challenging assembly projects owing to their size, multiple recombinogenic regions, and frequent transfer of DNA from the chloroplast. Assembling the mitochondrial genome for *Eleusine indica* (goosegrass) provides baseline genomic data for an economically significant invasive species that is also the maternal parent of the allotetraploid crop African finger millet (*Eleusine coracana).* The assembled genome is the product of various-length libraries: it contains 33 protein coding genes, 6 rRNA subunits, 24 tRNA, and 8 large repetitive regions, as well as 15 kp of transposable elements across a total of 520,691 bp. The *E. indica* mitochondrial genome shows evidence of RNA editing and has lost *rpl2*, *rpl5, rps14*, *rps11*, *sdh4,* and *sdh3* genes, consistent with a pattern that is widespread among Poaceae. Examination of additional mitochondrial coding sequences from Poaceae and an *in silico* Southern blot approach show high sequence conservation and provide new data on mitochondrial gene loss patterns within Poaceae.

Introduction

Plants contain three genomes: a large nuclear genome and two organellar genomes, mitochondrial (mt) and plastid (Leaver and Gray 1982; Gray 1989; Gray and Archibald 2012). The organellar genomes, the product of separate endosymbiotic events, have both retained the

standard genetic code and normally occur in quadripartite (plastid) or multipartite (mt) conformations (Palmer 1985; Gray 1989). Organellar genomes have transferred thousands of genes to the nucleus since the first endosymbiotic event, a process that is ongoing in plants where the rate of mt gene loss is rapid (Palmer and Herbon 1988; Adams et al. 2000; Palmer et al. 2000; Adams and Palmer 2003; Smith and Keeling 2015).

Among the three genomic compartments, the mt genome generally has the slowest rate of substitution and the fastest rate of rearrangement (Palmer and Herbon 1988). Gene shuffling within plant mt sequences is facilitated by recombination among linear segments of the genome and occurs so frequently that gene order can vary within individual organisms, rendering collinearity within plant mt genomes almost meaningless (Iorizzo et al. 2012). The recombination and multipartite nature make it possible to assemble the mt genome sequence into one or more circular chromosomes (Sloan 2013).

The assembly of plant mt genomes is not trivial and is outpaced by plastid genome assembly by approximately 7 to 1 according to NCBI Genomes (1,454 chloroplast genomes vs. 209 mt genomes sequenced https://www.ncbi.nlm.nih.gov/genome/browse/ Accessed: July 31, 2017). This imbalance is driven by the ease of plastid assembly which is increasingly solved through automated approaches and deeper genome sequencing (Soorni et al. 2017). In contrast, plant mt genomes represent a more difficult assembly problem for automated assembly that cannot be resolved by longer reads or greater sequencing depth alone. The correct assembly of a circularized plant mt genome currently requires manual curation because of properties intrinsic to the mt genome: first, repetitive recombinogenic regions with multiple and correct resolutions to the placement of recombinogenic DNA, and second, intracellular DNA transfer either from the chloroplast or, in some cases, horizontally transferred DNA from other organisms

(Bergthorsson et al. 2004; Sloan 2013; Soorni et al. 2017). The assembly of mt genomes remains an important and challenging task that provides substantive answers to questions such as mt gene content.

The variability of mt gene content provides insight into ongoing intracellular gene transfer within lineages of plants: there is an impressive range of gene content within plant mt genomes, ranging from as few as three genes to as many as 67 (Adams and Palmer 2003). The complete transfer of a gene to the nucleus involves multiple steps. First, a transfer of a functional gene copy to the nucleus which may require functional changes to the sequence to account for the lack of appropriate RNA editing in the nucleus. Second, the inserted sequence must acquire a promoter and a mt targeting sequence. Finally, the gene is lost from the mt genome (Brennicke et al. 1993; Adams and Palmer 2003; Bonen 2006). While ancient transfers eventually result in the elimination of the mt gene, there are many examples of recent transfers which are in transitional states. In these cases organisms either maintain a functional nuclear and mt gene, as in the case of *sdh4* and *Populus* (Choi et al. 2006)*, rpl5* and *Triticum aestivum* and other grasses (Sandoval et al. 2004; Wu et al. 2017), and *rps19* and *Bromus inermis* (Atluri et al. 2015), or they retain a mt pseudogene as in the case of *rps14* and *Arabidopsis* (Aubert et al. 1992) and in select grasses (Ong and Palmer 2006).

There are nine known genes lost within Poaceae: two are undergoing functional transfer within Poaceae *rpl2* (Subramanian and Bonen 2006) and *rpl5* (Sandoval et al. 2004; Wu et al. 2017), and seven were transferred prior to the emergence of Poaceae. Within the seven transfers, one occurred at the base of monocots *rpl10* (Adams et al. 2002; Mower and Bonen 2009; Kubo and Arimura 2010), two occurred in Poales *sdh3*, *sdh4* (Adams et al. 2001b), and four occurred within the common ancestor of Poaceae *rps10* (Adams et al. 2002)*, rps11* (Bergthorsson et al.

2004), *rps14* (Sandoval et al. 2004; Ong and Palmer 2006) and *rps19* (Fallahi et al. 2005; Atluri et al. 2015).

The progressive loss of a mt sequence may vary dramatically for any given gene; for example, *rps19* was transferred in the common ancestor of *Poaceae* yet its presence and function within the mt genome is variable. For example, it is present as a functional copy in *B. inermis*, as a pseudogene in *Triticum aestivum,* and is absent from *Hordeum vulgare* (Fallahi et al. 2005; Atluri et al. 2015).

In the absence of fully assembled genomes, mt gene loss is commonly inferred from Southern blot analysis (Palmer et al. 2000; Adams and Palmer 2003). Here, I develop an *in silico* analog to investigate gene content with Poaceae and demonstrate that it provides comparable results (Palmer et al. 2000; Adams et al. 2001a; Adams and Palmer 2003; Liu et al. 2009; Atluri et al. 2015; Wu et al. 2017).

Eleusine indica (L.) Gaertn. (Poaceae Chloridoideae) or goosegrass is a diploid (2n=18) member of the grass subtribe Cynodonteae and a widespread weed with multiple herbicide resistant strains (Mudge et al. 1984; Ng et al. 2003). It is a genome donor of, and frequently crosses with, allotetraploid *E. coracana* (L.) Gaertn. (African finger millet) (2n=36), a cereal crop grown on marginal soils that provides core nutritional support in many impoverished areas (Phillips 1972; Singh and Raghuvanshi 2012; Shobana et al. 2013; Hittalmani et al. 2017). Developing genomic resources for *E. indica* will improve understanding of a weedy grass species and increase the resources available for the improvement of an under-studied crop plant that provides critical nutrients in arid regions of East Africa and West India.

Methods

Sequence Preparation and Assembly

Whole genomic DNA was extracted from fresh leaves of *Eleusine indica* with DNeasy Plant Mini Kit (Qiagen, CA, USA) and used to create 2 paired-end and 1 mate-pair libraries (Table 1) that were sequenced using the Illumina platform as described in Zhang et al. (2016). Fastq files were filtered to exclude chloroplast or contaminating vector sequences with Bowtie 2 v2.2.9 (Langmead and Salzberg 2012) and SAMtools v1.2 (Li et al. 2009) and checked for quality and length with the perl pipeline Trim Galore v0.4.0 (Krueger 2015) and FastQC (Andrews and Others 2010) (Table 1).

A nearly complete pseudo-molecule of *ca*. 500 Kbp that contained 33 protein coding genes, 3 rRNA genes, and several recombinogenic repeats using an iterative assembly process was created, in which the output from multiple assemblers (Velvet v1.2.10 (Zerbino and Birney 2008)) with VelvetOptimizer v2.2.5 (Gladman and Seemann 2012), Ray v2.0.1 (Boisvert et al. 2010) and AllPaths-LG v3 (Gnerre et al. 2011) was compared to a reference mt genome of *Zea mays* (GenBank accession NC_007982) using blastn in conjunction with a series of python scripts (https://github.com/NDHall/EleusineMitochondria/tree/master/contig_viz). This preliminary assembly was used to identify unassembled read pairs for de novo assembly with Ray (Boisvert et al. 2010). Recombinogenic repeat regions were identified and joined using blastn, mapping depth, and the support of mate-pair and paired end reads (Mower et al. 2012). Residual gaps in the assembly were filled with contigs from previous shotgun assemblies. Assembly of chloroplast insertions were verified using mate-pair reads anchored in the chloroplast insertion and flanking mt sequence.

Validation

Final assembly was confirmed by even mapping depth of reads (filtered for edit distance of zero) across the entire reference using Bowtie 2 (Langmead and Salzberg 2012) and the connection of mate-pair and paired-end reads across all major junctions visualized as a connection plot created in R v3.4.0 (Team 2013). Figure 1 shows the position of all reads plotted against themselves and plotting the mate pair on the y-axis above (5' to 3' direction) or below (3' to 5') each read pair. Read depth was determined with SAMtools v1.2 (Li et al. 2009) from the map created in Bowtie 2 (end-to-end mapping and pairing enforced) for unfiltered mate-pair reads and cleaned paired-end reads (Table 1).

2.3.3 Annotation/Repetitive Fraction

Mitochondrial sequence was annotated with Mitofy (Alverson et al. 2010), NCBI BLAST+, and Artemis (Rutherford et al. 2000). Transfer RNAs were identified using tRNAScan and filtered for length less than 100 and COVE Score greater than 21. RepeatMasker version 3.2.7 (Smit et al. 1996), RepBase version 20090604 (Jurka et al. 2005), and 31 unique sequences with length of greater than 100 were annotated and extracted. These segments were compared to the Reference Mitochondrial Genomes of all angiosperms available on NCBI (Accessed October 2016). Tandem repeats were identified using Misa (http://pgrc.ipk-gatersleben.de/**misa:** Accessed 13 Dec 2016) with a minimum of 5 repetitions for mono and di nucleotide repeats, minimum of 3 repetitions for tri to octa nucleotide repeats, and minimum of 2 repetitions for nona to 31X nucleotide repeats.

SRA Consensus Assembly

Fastq files were mapped to chloroplast and mt sequence references using Bowtie 2 (Langmead and Salzberg 2012) local alignment with match bonus, qc-filter, and up to 900M reads. Bam files were filtered to exclude reads that mapped to chloroplast references, converted to fastq files, and processed with Trim Galore (Krueger 2015) using illumina flag (--illumina) which targets published adapters used for Illumina sequencing. Filtered reads were mapped back to the mt reference using the original parameters, resulting maps were filtered with cigar_filter.py with minimum match length of 25 (https://github.com/NDHall/pysam_tools/tree/master/cigar_filter), and then passed through the best practice Picard and GATK pipeline to the realignment step (Van der Auwera et al. 2013) using GATK v3.6 (McKenna et al. 2010). Realigned bam files were used to produce reference-free consensus sequences using pysam_consensus.py and minimum depth of 6 (https://github.com/NDHall/pysam_tools/tree/master/consensus_caller). A modified version of fasta-stats.py (https://techoverflow.net/2013/10/24/a-simple-tool-for-fasta-statistics/ Accessed July 20, 2017) was used to produce input for fasta_stats_parser.py (https://github.com/NDHall/pysam_tools/tree/master/fasta_stats) with default settings which limited sequences to high-quality, high-coverage consensus sequences for broader Poaceae to produce a list of acceptable assemblies which were extracted with select_contigs.pl (https://github.com/chrishah/phylog/blob/master/scripts-external/select_contigs.pl) and aligned in Mafft v7.123 (Katoh and Standley 2013). Alignments were visually checked and manually curated to be inframe using SeaView (Gouy et al. 2010). Taxa were retained for further analyses if they had more than than 10,000 nucleotide positions out of 27,848 possible, and each alignment was ordered and expanded to include missing taxa using fasta_ghost.py (https://github.com/NDHall/pysam_tools/tree/master/fasta_ghost) and concatenated using

FASconCAT (Kück and Meusemann 2010). Final sequences were spot checked against known reference genomes using Blast search utility on NCBI.

RNA editing

Monoisolate transcriptome reads were downloaded from NCBI SRA (GenBank Accesion ERR1590130), checked with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed for quality with Trimmomatic v.0.3.3 (Bolger and Giorgi 2014). Quality trimming was confirmed with FastQC (Krueger 2015). Two mappings were used to locate potential RNA edits. To search for putative non-coding edits, the finished mt genome and plastid genome (GenBank Accession NC_030486) were made into a single reference, and to search within coding regions, coding sequences were extracted from mt genome (GenBank Accession MF616338) and plastid genome (GenBank Accession KU833246) genomes and made into a combined reference. For both non-coding and coding reference files, reads were mapped with Bowtie 2 (Langmead and Salzberg 2012) employing local alignment and match bonus. For genomic mapping and end to end mapping for coding sequences, duplicates were marked with Picard Tools and variants were called with BCF tools v1.5 (https://github.com/samtools/bcftools). C-U edits were counted for coding regions by tallying all C-T variants within mt CDS sequences, and for non-coding regions C-U edits were tallied C-T or T-C variants on the mt genome after the exclusion of coding regions using vcftools v0.1.14-14 (Danecek et al. 2011). Minimum accepted variant depth was set at frequency of 0.2. Coding sequences for the 33 protein coding genes were extracted from the annotation and 1 copy of each gene was submitted to Prep-Mt with a cutoff of 0.2 (Mower 2005).

Phylogenetic Analysis

A full codon by gene partition scheme was examined with Partition-Finder v2.0.0 (Lanfear et al. 2012), model selection was limited to GTR-GAMMA and GTR-GAMMA+I with greedy search algorithm, and the best scheme was used for subsequent phylogenetic analysis. Trees were created using RAxML-MPI-AVX v8.2.6 (Stamatakis 2014) with 100 rapid bootstraps, GTRGAMMA model, and the best partition scheme returned by Partition-Finder. Trees were visualized with FigTree (Rambaut 2009).

Individual rates of gene evolution for chloroid samples, *Eleusine indica* (Auburn, Table 1), *Sporobolus michauxianus* (GenBank Accession SRR556090), and *Oropetium thomaeum* (SRR2083764) were determined as simple counts of nucleotide substitutions made by Mega-cc (Kumar et al. 2012) from nucleotide alignments created with Clustal (Thompson et al. 1994). To make the comparison among subunits as even as possible, marginal consensus sequences were added where available, and all sites containing gaps were excluded from the analysis and rates of substitution were compared among subunits. Chlorid samples were analyzed with Codeml from PAML 4.9b (Yang 2007) which was used to determine dN/dS ratios for subunits with high substitution rates, with *Ananas comosus* (GenBank Accession DRR022930) as the outgroup. To put these observations in a larger phylogenetic context, all available high quality sequences were analyzed using Codeml in PAML v4.9b (Yang 2007). A reference tree based on GWGP II (Grass Phylogeny Working Group II 2012) was used to quantify rates of exon evolution for each taxa. Sequences were taken from consensus sequences prepared as before and gene alignments were created with codon aware Mega-cc (Kumar et al. 2012) visualized and curated with SeaView (Gouy et al. 2010). Trees and alignments were prepared using custom python scripts which

produced a reduced tree for each gene set based on the number of taxa included in it and produced a sequential PAML style phylip file, in which taxon labels were terminated by 2 spaces. Codeml was run for both chloroid and Poaceae sets with free rates, ambiguous sites removed, a starting kappa of 2.0 and a starting omega of 0.4. Resulting dN and dS trees were printed to a standard size using FigTree (Rambaut 2009)

Heat Map

Whole genomic or unspecified SRA Poaceae datasets were downloaded (Supplementary Table 1) and passed through Heat Map Pipeline (Supplementary Fig. 1) with a BEP or PACMAD clade-specific reference. Each SRA readset was mapped to this reference which contained both target mt sequence and chloroplast genomes used to exclude unwanted reads. A species-specific consensus sequence for each mt gene was created from the initial map and a second mapping of the original reads was performed against this consensus. The read depth observed for each gene was normalized by dividing the average combined depth of the three longest, highly conserved regions: *matR*, *nad4* and *nad7*. Exons and introns were mapped separately so that reads overlapping the intron exon boundary were not lost: this was of particular concern for lower coverage sets. Low coverage duplicates or inconsistent coverage read sets were discarded. Genes were determined to be sufficiently covered for representation in the heat map if average raw read depth was greater than 5 for more than one-half of the sequence.

Heat map color was assigned based on sequence coverage relative to the average depth of mapping across the three control genes. Sequences that are absent are black, while ultra-high-

coverage chloroplast sequences are presented as light yellow or white. The fold mitochondrial depth was transformed with the natural log and plotted across the length of the gene for a control region *matR* and variable regions, *rpl2*, *rpl5*, *rpl10ψ*, *rps11*, *rps14*, and *rps19* using Python 3.4.3 with the matplotlib (Hunter 2007) and pandas (McKinney 2010) libraries. To determine if the *rpl5* and *rps14* genes were undergoing tandem loss, the above pipeline and graphing was also performed with selected taxa [*Alloteropsis cimicina* (GenBank Accession SRR2163548), *Aristida congesta* (GenBank Accession: SRR2163568), *Aristida purpurea* (GenBank Accession: SRR2163569), *Danthoniopsis dinteri* (GenBank Accession: SRR2163566), *Echinochloa frumentacea* (GenBank Accession: SRR2162759), *Eleusine indica*, *Eragrostis tef* (GenBank Accession: SRR1463402), *Leersia perrieri* (GenBank Accession: SRR1528439), *Oropetium thomaeum* (GenBank Accession: SRR2083764), *Oryza longistaminata* (GenBank Accession: SRR1264538), *Oryza punctata* (GenBank Accession: SRR1264539), *Sporobolus michauxianus* (GenBank Accession: SRR486071), *Triticum monococcum* (GenBank Accession: SRR445609), *Triticum turgidum* (GenBank Accession: ERR463920), *Triticum uratu* (GenBank Accession: ERR424867)] for a 10,044 bp region of *Oryza sativa (*GenBank Accession: NC_007886.1:340483-350527) containing 3' end of *cox1*, intergenic spacers, *rps14*, and *rpl5,* using the *Zea mays matR, nad7* and *nad4* sequences as controls for depth normalization.

Results

Sequence Preparation and Validation

Sequencing of the three DNA libraries produced 202 million reads which was reduced to 178 million through the process of read filtering and quality control (Table 1). The final master

circle is 520,691 bp (GenBank accession MF616338) and comprises 3% of the 178 million filtered reads.

The evenness of read pairing and depth indicate that both read-pair and mate-pair data support the master circle assembly across multiple recombinogenic regions (Fig. 1a). Discordant pairing creates areas of low coverage in recombinogenic regions and makes these regions visible in the connection plot (Fig. 1b). Slightly increased coverage in chloroplast transfer regions is an artifact of plastid reads included in the validation mapping step (Table 1).

2.4.2 Annotation/Repetitive Fraction

A total of 33 mt protein coding genes, 2 pseudogenes (*rpl2*, *rpl10*), 6 rRNA (2 copies each of 3 subunits), and 24 total tRNA with 15 unique tRNAs (Fig. 2) were annotated. Analysis with tRNAScan also identified a large, 120bp, putative tRNA which was excluded because it is a pseudogenized sequence (Gualberto et al. 1988; Wintz et al. 1988). Two chimeric ORFs that encode an *atp8*, plus attached domains of either ELF or DUF, were identified. Twelve different plastid transfers greater than 400 bp were detected that contained multiple ORFs containing fragments of *psaA*, *psaB*, *rpl2-ct*, *Ycf2*, *atpA*, *ndhJ*, and 9 tRNA of apparent plastid origin. All tRNA identified by COVE implemented with Mitofy (Alverson et al. 2010), with the exception of *trnC,* occur in or are closely associated with these larger regions of transferred chloroplast sequence. The *trnC* occurs within a possible insertion of 150 bp that includes the *trnC* and flanking intergenic spacers from the *Eleusine indica* plastid genome identified using BLAST+.

Eight large repetitive regions greater than 400 bp were identified with blastn and designated A (11461bp), B (8059bp), C (6378bp), D (2889bp), E (1968bp), F (1042bp), G (527), and H (417bp) (Fig. 1). Three of the repeat regions contained genes and retro-elements: Repeat A contains 1 rRNA (*rrnL*), 1 Copia element, and a *rpl2* pseudogene; Repeat B contains 2 LTR

elements, 2 rRNA (*rrn5*, *rrnS*), and the *cox1* gene; and Repeat C contains 1 Copia element (Fig. 1). The other five repeats appeared to be purely recombinogenic sequences with no detected features.

RepeatMasker identified several transposable elements within the mt genome that covered more than 15 kbp (2.9% of total mt genome); these elements exhibit a range of occurrences across angiosperm mt genomes. Retrotransposons were a majority of the identified elements (14,553 bp, 3% mt genome) split unevenly between LTR retroelements which contain Gypsy/DIRS1 (7807 bp, 1.5% mt genome) and TY1/Copia (6368 bp, 1.2% mt genome) as well as LINEs which contain L1/CIN4 (378 bp, 0.073% mt genome). DNA transposons accounted for 862 bp (0.17% MtG) of sequence among the families Tourist/Harbinger 457 bp (0.088% MtG), EN-Spm 296 bp (0.057% MtG), and MuDR-IS905 109 bp (0.021% MtG).

RepeatMasker (Smit et al. 1996) identified a total of 31 unique elements greater than 99 bp. Nine of the elements are unique to the mitochondria of *Eleusine indica*, 20 are known from monocots, and 2 occur throughout angiosperms. *Copia1-VV_I* occurs throughout angiosperms and was identified in monocots and eudicots including the Asterid *Capsicum annuum* L. (Solanaceae). *Monkey_MA* occurs throughout angiosperms, though the overlap between it and *Copia1-VV_I* is minimal. Misa identified a total of 3209 SSRs, a majority of which were mono-nucleotide repeats (2820) and nearly a third of the repeats (1630) were in compound formation (Table 2).

RNA Editing

Within the coding region, Prep-Mt (Mower 2005) predicted a total of 426 RNA putative edit sites in the coding sequences, with TCA to TTA the most common (occurring 66 times). The five least common occurred only once, ACC to ATC, CAA to TAA, CAG to TAG, and CGA to TGA. Serine to Leucine and Proline to Leucine were the two most frequent amino acid changes, occurring 102 and 82 times, respectively. When results of Mt-Prep were combined with read mapping predictions, a total of 620 putative RNA editing sites were identified, of which 530 were in coding regions and 430 were well supported (Table 3). At the level of individual loci, *ccmFn* (37) and *ccmC* (36) had the highest number of well-supported edits and *nad3* (0) and *eleind009* (0) showed an absence of well-supported edit sites (Table 4). Within non-coding regions, 90 edits were identified of which 6 were well-supported at sites with mapping greater than 10.

2.4.4 Phylogenetic Analysis

One hundred and fourteen SRA datasets were downloaded and, when filtered and mapped to reference exons, 3.3M reads were kept for processing in consensus pipeline. After consensus calling, alignment, and filtering, 71 datasets were used to create the final supermatrix which was 74.08% complete and had 3,138 distinct alignment patterns.

RAxML (Stamatakis 2014) consistently recovered the chloroid clade and correctly placed *Eleusine* sister to *Oropetieum* and *Sporobolus* sister to *Eleusine* and *Oropetium*. Internal nodes within the tree were often poorly supported and incongruent with accepted relationships, with the *Oryza/Leersia* clade was incorrectly placed and there was poor resolution for the undersampled portions of the PACMAD clade (Supplementary Fig. 2).

The atpase subunit showed an increased rate of nucleotide substitution within chloroids (Table 5). Statistical analysis with codeml confirmed that substitution rate for atpase was elevated and that it was driven by synonymous substitutions (Table 6, Supp. Fig. 3).

Heat map and depth analysis

As stated previously, according to Bergthorsson et al. (2003), Adams and Palmer (2003) Sandoval et al. (2004), Fallahi et al. (2005), Subramanian and Bonen (2006), Kubo and Arimura (2010), and Atluri et al. (2015), there are several known functional gene transfers prior to the origin of the Poaceae (*rps10*, *rpl10*, *rps11*, *rps14*, *rps19, sdh3* and *sdh4)* and two gene transfers that are occurring within the family (*rpl2* and *rpl5*). Testing for the presence of these variable genes reveals the infra-familial pattern of mt gene transfer and mt sequence loss since a functional transfer to the nucleus is not concurrent with loss of mt sequence. Furthermore, surveys of mt gene presences have the potential to highlight rare cases of gene regain via introgression sequence (Brennicke et al. 1993; Adams and Palmer 2003; Bonen 2006). Using available SRA data, I examined the presence/absence of mt genes from a collection of 70 species of Poaceae and one species of Bromeliaceae by *in silico* analysis of read depth (Figure 3, Supp. Fig. 4). The resulting heat map showed diverse patterns of loss over time in Poaceae. Most genes, particularly those involved in oxidative phosphorylation, have been retained in all lineages, consistent with previous studies (Adams et al. 2002; Adams and Palmer 2003). In contrast, several genes (*rps10*, *sdh3* and *sdh4)*) are absent from all Poaceae species, indicating that they were lost prior to the origin of grasses. Six ribosomal protein genes (*rpl2*, *rpl5*, *rpl10*, *rps11*, *rps14* and *rps19*) exhibited a varied pattern of differential retention and loss during Poaceae evolution.

To assess the accuracy of the heat map analysis in determining the presence or absence of genes in the mt genome, I compared the heat map results to the sequenced *Eleusine indica* mt genome, and to eight other Poaceae mt genomes available in GenBank (*Aegilops speltoides*, *Hordeum vulgare*, *Lolium perenne*, *Oryza sativa*, *Sorghum bicolor*, *Tripsacum datcyloides*, *Zea mays* subsp. *mays*, and *Zea mays* subsp. *parviglumis*). The heat map confirmed that all nine species lacked the *rps10*, *sdh3*, and *sdh4* genes, as expected. Among the six variably present ribosomal protein genes, the heat map confirmed that *Sorghum*, *Tripsacum*, and *Zea* lost all six genes, *Eleusine indica* retained *rps19*, *Hordeum* retained *rpl5*, *Lolium* retained *rpl5* and a large *rps14* pseudogene, *Aegilops* retained *rpl5* and two large pseudogenes (*rps14* and *rps19*), and *Oryza* retained *rpl2*, *rpl5*, *rps19*, and three large pseudogenes (*rpl10*, *rps11*, *rps14*). The heat map analysis failed to detect several smaller (<100 bp) pseudogenes of *rpl10* and *rps19* in a few species, suggesting a rough limit of detection for this approach (Fig. 3).

Discussion

General features of *Eleusine* mitochondrial genome

Discordant mate-pair mapping caused by recombinogenic activity served to highlight the difficulty posed by this phenomenon for assembling plant mt genomes, many of which have been intransigent to automated assembly techniques employing long-read sequencing technologies (Soorni et al. 2017). As is the case with many Poaceae mt genomes, the *Eleusine* mt genome

exhibited a typical set of mt genes, with foreign DNA from the plastid and nucleus and several likely recombinogenic repeats that promoted a multipartite structure.

Analysis of putative RNA editing sites yielded a range 436-620 sites generally in line with published grass genomes (Mower 2009). The lack of phylogenetic resolution is not surprising given the sequence conservation (Palmer and Herbon 1988). Investigation of the substitution rates for individual genes showed an apparent pattern of slow rates for *nad* genes and accelerated rates for *atp* and *cox* genes which followed the broad pattern of substitutions reported by Cui et al. (2009).

Heat map and depth analysis

The heat map analysis, coupled with read depth and inspection of assembled gene sequences, suggested frequent loss of coding regions, particularly associated with, *rps19, rpl2* and *rpl5*, across Poaceae (summarized in Fig. 4). The results of depth plotting, by gene, heat map and sequence assembly showed broad patterns of loss over time. The depth plots showed the successive loss of r*pl2* sequence within PACMAD clade and Pooideae in the BEP clade, as different portions of the pseudogenized sequence were lost over time (Supplementary Fig. 4). The mt sequence for *rpl5* was well characterized in the both PACMAD and BEP clades of Poaceae. Additional loss of *rpl5* from the mt genomes of *Setaria, Cenchrus, Eleusine,* and *Panicum* are evident. The assembled consensus sequences for *rpl5* strongly suggested pseudogenization in *Sporobolus* and *Echinochloa,* and the likelihood a functional copy within *Alloteropsis* was retained. The variable loss of *rpl5* from the mitochondria within Paniceae was striking, since it appeared that both *Alloteropsis* and *Echinochloa* retained mt copies of the gene, while *Oplismenus* lost its copy. The pseudogene *rps14* showed a similar pattern as *rpl5* within

Alloteropsis and *Echinochloa.* This was likely related to the fate of *rpl5* situated directly upstream of *rps14*. The loss of *rps19* from within the BEP portion of Poaceae has been studied extensively (Atluri et al. 2015). The results concured with those previously reported, and expand knowledge of *rps19* loss and retention within PACMAD clade. Plotting the depth of the genes showed that within the PACMAD clade only Andropogoneae, Arundinelleae and Paspaleae showed widespread loss of its copy of the *rps19* sequence. Consensus calling across maps of *rps19* strongly suggested it persisted as a functional copy within Aristidoideae, Chloridoideae, and Paniceae. Within PACMAD clades possessing *rps19*, its loss remained variable. For example, within Paniceae, *Cenchrus* showed evidence of having completely lost *rps19* from its mt genome, the first reported loss for *rps19* within Paniceae. The *rps11* pseudogene was lost across PACMAD clade, and within Triticeae and Poeae 2 clades of BEP. Furthermore, there appeared to be a recent, novel loss of *rps11* sequence within *Oryza brachyantha,* unique within that economically important genus.

Our analysis also increased the depth of the *rpl2* transfer within PACMAD clade. Examination of read depths across gene length within *rpl2* containing PACMAD taxa showed a range of patterns from the most complete samples which exhibit a deletion in exon 1 and in the group II intron (Subramanian and Bonen 2006) to the functional absence of the any *rpl2* sequence (Supplementary Fig. 4). Assembly of a representative deeply sequenced sample, *Oropetium thomaeum* (GenBank accession SRR2083764), produced a single contig with structural similarity to the *rpl2* pseudogene from *Bromus inermis* (GenBank accession KT022083.1). The apparent pseudogenization of *rpl2* in all surveyed PACMAD sequences suggested a common transfer of *rpl2* at the base of PACMAD clade. Similar patterns of loss were evident for Poaeae-2 and Triticeae consistent with previous reports (Subramanian and Bonen 2006)

Visualizing copy number of putative mt sequences *in silico* is a potentially broadly useful tool for plant mt genome analysis. The results for gene content closely matched those reported within the literature (e.g., Alturi et al. 2015; Wu et al. 2017), with a few discrepancies that provided further insight into the gene content and evolutionary dynamics within mt genomes of Poaceae. There are apparent cases of ongoing losses within species such as *Danthoniopsis dinteri*, in which a pseudogene for *rpl5* was reported (Wu et al. 2017 Supplementary Material) and a pseudo-gene was reported for *rps14* (Ong and Palmer 2006), but the sampling of SRA data (GenBank Accession: SRR2163566) suggested that both *rpl5* and *rps14* were missing from the individuals sampled (Figs. 4, 5 and Supplementary Fig. 2). Wu et al. (2017) reported *rpl5* is likely missing from *Triticum urartu*: my results concured and showed that it is indeed absent from the mt genome by virtue of its low depth of coverage (GenBank Accession: ERR424867). Also within *Triticum, T. monococcum* (GenBank Accession: SRR445609) was missing *rpl5,* which was retained or regained within *T. turgidum* (GenBank Accession: ERR463920). *Triticum uratu* and *T. monococcum* are diploid (AA) genomes and *T. turgidum* is a tetraploid (AABB) genome, so a potential scenario is that in which *rpl5* and *rps14* genes (Fig. 4, Fig. 5) were lost from the mt genomes of *T. uratu* and *T. monococcum* and restored during the allopolyploidization event that created *T. turgidum.* The genes *rpl5* and *rps14* are tightly linked within virtually all angiosperms (Ong and Palmer 2006). As a consequence of their proximity, the fates of r*pl5* and *rps14* were frequently linked (Figs. 4, 5). Notable exceptions were *Sporobolus michauxianus* (GenBank Accession: SRR486071) and *Leersia perrieri* (GenBank Accession: SRR1528439) in which *rps14* was lost independently of *rpl5.* The loss of *rps14* in the context of its ancestral spatial relationship relative to *rpl5* provided further insight into the rate and scale of sequence loss within plant mt genomes. Two other pseudogenized sequences, *rps11*

and *rpl10,* were sparsely reported on within Poaceae but this approach provided insight into the heterogeneous loss of mt sequence. The *rpl10* and *rps11* pseudogenes are generally present within the BEP clade and absent from the PACMAD clade, but there are a few exceptions to this broad pattern that were found within my sampling. The loss of *rps11* in *Oryza brachyantha* (GenBank Accession: SRR350709) appeared to have happened after the divergence of the FF *Oryza* genomes from the rest of *Oryza* (Nishikawa 2005). The *rpl10* pseudogene was present in *Aristida* yet absent throughout the rest of the PACMAD clade samples, which suggested it was lost after the divergence of *Aristida* from the rest of the PACMAD clade. Finally, these data provided deeper insight into the functional transfers of *rpl2* within the PACMAD Poaceae. It is intriguing that the *rpl2* nuclear transfers were from different events, yet the pseudogene sequences shared strong similarity to the *Bromus inermis* pseudogene which had deletions in exon 1 and the group II intron (Subramanian and Bonen 2006).

A broad overview of mt genome gene loss throughout Poaceae was presented by gleaning information from several whole genomic datasets for grasses of variable coverage with no published mt genomic sequence. The *in silico* heat map approach allowed gene loss determination without full assembly, and it made possible the survey of loci that are rarely reported on such as *rpl10* and *rps11* with little extra effort compared to traditional methods. The *Eleusine indica* mt genome and the broader context of gene loss in Poaceae represents a basic step toward a fully genomic understanding of a prolific herbicide resistant weed and maternal progenitor of an important allotetraploid crop.

References

Adams KL, Daley DO, Qiu YL, et al. (2000) Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. Nature 408:354–357

Adams KL, Ong HC, Palmer JD (2001a) Mitochondrial gene transfer in pieces: fission of the ribosomal protein gene *rpl2* and partial or complete gene transfer to the nucleus. Mol Biol Evol 18:2289–2297

Adams KL, Palmer JD (2003) Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phylogenet Evol 29:380–395

Adams KL, Qiu Y-L, Stoutemyer M, Palmer JD (2002) Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci U S A 99:9905–9912

Adams KL, Rosenblueth M, Qiu YL, Palmer JD (2001b) Multiple losses and transfers to the nucleus of two mitochondrial succinate dehydrogenase genes during angiosperm evolution. Genetics 158:1289–1300

Alverson AJ, Wei X, Rice DW, et al. (2010) Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol Biol Evol 27:1436–1448

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

Atluri S, Rampersad SN, Bonen L (2015) Retention of functional genes for S19 ribosomal protein in both the mitochondrion and nucleus for over 60 million years. Mol Genet Genomics 290:2325–2333

Aubert D, Bisanz-Seyer C, Herzog M (1992) Mitochondrial *rps14* is a transcribed and edited pseudogene in *Arabidopsis thaliana*. Plant Mol Biol 20:1169–1174

Bergthorsson U, Adams KL, Thomason B, Palmer JD (2003) Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature 424:197–201

Bergthorsson U, Richardson AO, Young GJ, et al. (2004) Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm *Amborella*. Proc Natl Acad Sci U S A 101:17747–17752

Boisvert S, Laviolette F, Corbeil J (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol 17:1519–1533

Bolger A, Giorgi F (2014) Trimmomatic: a flexible read trimming tool for illumina NGS data. URL http://www usadellab org/cms/index php

Bonen L (2006) Mitochondrial genes leave home. New Phytol 172:379–381

Brennicke A, Grohmann L, Hiesel R, et al. (1993) The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants. FEBS Lett 325:140–145

Choi C, Liu Z, Adams KL (2006) Evolutionary transfers of mitochondrial genes to the nucleus in the *Populus* lineage and coexpression of nuclear and mitochondrial *Sdh4* genes. New Phytol 172:429–439

C J Leaver, Gray MW (1982) Mitochondrial genome organization and expression in higher plants. Annu Rev Plant Physiol 33:373–402

Danecek P, Auton A, Abecasis G, et al. (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Fallahi M, Crosthwait J, Calixte S, Bonen L (2005) Fate of mitochondrially located S19 ribosomal protein genes after transfer of a functional copy to the nucleus in cereals. Mol Genet Genomics 273:76–83

Gladman S, Seemann T (2012) VelvetOptimiser. Victorian Bioinformatics Consortium, Clayton, Australia: http://bioinformatics net au/software velvetoptimiser shtml

Gnerre S, Maccallum I, Przybylski D, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 108:1513–1518

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221–224

Grass Phylogeny Working Group II (2012) New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. New Phytol 193:304–312

Gray MW (1989) Origin and evolution of mitochondrial DNA. Annu Rev Cell Biol 5:25–50

Gray MW, Archibald JM (2012) Origins of mitochondria and plastids. In: Bock R, Knoop V (eds) Genomics of Chloroplasts and Mitochondria. Springer Netherlands, pp 1–30

Gualberto JM, Wintz H, Weil J-H, Grienenberger J-M (1988) The genes coding for subunit 3 of NADH dehydrogenase and for ribosomal protein S12 are present in the wheat and maize mitochondrial genomes and are co-transcribed. Mol Gen Genet 215:118–127

Hittalmani S, Mahesh HB, Shirke MD, et al. (2017) Genome and transcriptome sequence of Finger millet (*Eleusine coracana* (L.) Gaertn.) provides insights into drought tolerance and nutraceutical properties. BMC Genomics 18:465

Hunter JD (2007) Matplotlib: A 2D graphics environment. Comput Sci Eng 9:90–95

Iorizzo M, Senalik D, Szklarczyk M, et al. (2012) De novo assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence

of DNA transfer into an angiosperm plastid genome. BMC Plant Biol 12:61

Jurka J, Kapitonov VV, Pavlicek A, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780

Krueger F (2015) Trim Galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Kubo N, Arimura S-I (2010) Discovery of the *rpl10* gene in diverse plant mitochondrial genomes and its probable replacement by the nuclear gene for chloroplast *RPL10* in two lineages of Angiosperms. DNA Res 17:1–9

Kück P, Meusemann K (2010) FASconCAT: Convenient handling of data matrices. Mol Phylogenet Evol 56:1115–1118

Kumar S, Stecher G, Peterson D, Tamura K (2012) MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics 28:2685–2686

Lanfear R, Calcott B, Ho S, Guindon S (2012) Partition-Finder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution 29: 1695–1701

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359

Li H, Handsaker B, Wysoker A, et al. (2009) The Sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Liu S-L, Zhuang Y, Zhang P, Adams KL (2009) Comparative analysis of structural diversity and sequence evolution in plant mitochondrial genes transferred to the nucleus. Mol Biol Evol 26:875–891

McKenna A, Hanna M, Banks E, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

McKinney W (2010) Data structures for statistical computing in Python. Proc. of the 9th Python in Science Conference

Mower JP (2005) PREP-Mt: predictive RNA editor for plant mitochondrial genes. BMC Bioinformatics 6:96

Mower JP (2009) The PREP suite: predictive RNA editors for plant mitochondrial genes,

chloroplast genes and user-defined alignments. Nucleic Acids Res 37:W253–9

Mower JP, Bonen L (2009) Ribosomal protein L10 is encoded in the mitochondrial genome of many land plants and green algae. BMC Evol Biol 9:265

Mower JP, Case AL, Floro ER, Willis JH (2012) Evidence against equimolarity of large repeat arrangements and a predominant master circle structure of the mitochondrial genome from a monkeyflower (*Mimulus guttatus*) lineage with cryptic CMS. Genome Biol Evol 4:670–686

Mudge LC, Gossett BJ, Murphy TR (1984) Resistance of goosegrass (*Eleusine indica*) to dinitroaniline herbicides. Weed Sci 32:591–594

Ng CH, Wickneswari R, Salmijah S, et al. (2003) Gene polymorphisms in glyphosate-resistant and-susceptible biotypes of *Eleusine indica* from Malaysia. Weed Res 43:108–115

Ong HC, Palmer JD (2006) Pervasive survival of expressed mitochondrial *rps14* pseudogenes in grasses and their relatives for 80 million years following three functional transfers to the nucleus. BMC Evol Biol 6:55

Palmer JD (1985) Comparative organization of chloroplast genomes. Annu Rev Genet 19:325–354

Palmer JD, Adams KL, Cho Y, et al. (2000) Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc Natl Acad Sci U S A 97:6960–6966

Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. J Mol Evol 28:87–97

Phillips SM (1972) A survey of the genus *Eleusine* Gaertn. (Gramineae) in Africa. Kew Bull 27:251–270

Rambaut A (2009) FigTree. Tree figure drawing tool version 1.3. 1. Institute of Evolutionary biology, University of Edinburgh: http://tree.bio.ed.ac.uk/software/figtree/

Rutherford K, Parkhill J, Crook J, et al (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945

Sandoval P, León G, Gómez I, et al (2004) Transfer of *RPS14* and *RPL5* from the mitochondrion to the nucleus in grasses. Gene 324:139–147

Shobana S, Krishnaswamy K, Sudha V, et al (2013) Finger millet (Ragi, *Eleusine coracana* L.): a review of its nutritional properties, processing, and plausible health benefits. Adv Food Nutr Res 69:1–39

Singh P, Raghuvanshi RS (2012) Finger millet for food and nutritional security. Afr J Food Sci 6:77–84

Sloan DB (2013) One ring to rule them all? Genome sequencing provides new insights into the "master circle"model of plant mitochondrial DNA structure. New Phytol 200:978–985

Smit AFA, Hubley R, Green P (1996) 2010 RepeatMasker Open-3.0. URL: http://www repeatmasker org

Smith DR, Keeling PJ (2015) Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. Proc Natl Acad Sci U S A 112:10177–10184

Soorni A, Haak D, Zaitlin D, Bombarely A (2017) Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. BMC Genomics 18:49

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Subramanian S, Bonen L (2006) Rapid evolution in sequence and length of the nuclear-located gene for mitochondrial L2 ribosomal protein in cereals. Genome 49:275–281

Team R (2013) R development core team. RA Lang Environ Stat Comput 55:275–286

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Van der Auwera GA, Carneiro MO, Hartl C, et al. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11.10.1–33

Wintz H, Grienenberger JM, Weil JH, Lonsdale DM (1988) Location and nucleotide sequence of two tRNA genes and a tRNA pseudo-gene in the maize mitochondrial genome: evidence for the transcription of a chloroplast gene in mitochondria. Curr Genet 13:247–254

Wu Z, Sloan DB, Brown CW, et al. (2017) Mitochondrial retroprocessing promoted functional transfers of *rpl5* to the nucleus in grasses. Mol Biol Evol 34:2340–2354

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

Zhang H, Hall N, Scott McElroy J, et al. (2016) Complete plastid genome sequence of goosegrass (*Eleusine indica*) and comparison with other Poaceae. Gene. doi: 10.1016/j.gene.2016.11.038

Table 1 Read set summary statistics

Summary	**Facility**	**Insert size (bp)**	**Fastq Stats**			**Trim galore flags**	**Filter Sequences**	**Mean Quality Scores**	**Mapped Reads**		**Mapping Depth**
			Raw Count	*Filtered Count*	*Read Length After QC (bp)*				*Total*	*Edit distance of 0*	*mean*
	Auburn	200	91,460,774	75,619,130	75-100	--paired --adapter2 TTCTTTCC CCCACCCT TTCC --length 75	KU666001.1	36.44	3,317,708	2,870,544	5571
	GeneWiz	400	100,277,890	82,788,746	90-101	--paired --illumina --length 90	NC_030486	36.92	2,139,526	2,031,907	416
	Genewiz	7.0e4	110,263,211	21,181,365	90-101	--paired --illumina --nextera	NC_030486	37.09	218,616	20,1672	40
Total Paired End			191,738,664	158,407,876							
Total combined			202,001,875	179,589,241							

* Longer version (MTV7) of sequence was used to include more reads, due the circular nature of the sequence this allows reads to mapped that would rwise be cut off.

Table 2 Summary of Misa results

Unit Size	No. SSRs
1	2820
2	21
3	270
4	53
5	16
6	3
9	9
10	3
11	1
12	3
14	1
15	2
17	1
20	1
22	1
24	1
26	1
27	1
30	1

Table 3 Total unique calls. List of non overlapping C to T calls made with alignment of extracted coding sequences mt-prep and bcf calls, allowing for a minimum depth of 1 and a minimum alternate allele frequency of 0.20. Well supported alleles include, all predictions shared between mt-prep and bcf calls, all bcf calls with a depth greater than 10 and a minimum alternate allele frequency of 0.20 and all unique mt-prep calls unsupported by mapping if mapping depth was less than or equal to 10. *The gene eleind009 was not submitted to mt-prep for analysis.

Sites	**Depth categories**	**total count**	**counts**
all unique calls		530	
	d > 10		87
	0 < d <= 10		373
	d = 0		70
all mt-prep		426	426
	d > 10		70
	0 < d <= 10		286
	d = 0		70
all bcf calls		380	
	d > 10		75
	0 < d <= 10		305
	d = 0		0
all shared		276	
	d > 10		58
	0 < d <= 10		218
	d = 0		0
unique bcf		104	
	d > 10		17
	0 < d <= 10		87
	d = 0		0
unique mt prep		150	
	d > 10		12
	0 < d <= 10		68
	d = 0		70
total well supported		**431**	

Table 4 Well-supported RNA editing calls compared with all calls, the total set of unique calls from Table 3. Calls are compared by gene and codon position. Average depth calculated by dividing sum of depth for all mapped positions by length of the gene. *Not submitted to mt-prep for analysis.

Category	avg. depth	counts per gene		codon	counts per codon	
		well supp.	all		well supp.	all
atp1	26.78	6	7			
				2	4	5
				3	2	2
atp4	4.59	9	10			
				1	2	2
				2	7	8
atp6	4.47	14	17			
				1	4	4
				2	10	10
				3		3
atp8	14.32	6	6			
				1	3	3
				2	2	5
				3	1	2
atp9	3117.54	7	7			
				1	1	1
				2	6	6
ccmB	0.24	31	31			
				1	12	12
				2	19	19
ccmC	3.10	36	40			
				1	16	17
				2	20	20
				3		3
ccmFc	0.99	21	27			
				1	10	12
				2	11	11
				3		4
ccmFn	90.07	37	41			
				1	15	16
				2	19	22
				3	3	3
cob	4.69	15	16			
				1	9	9
				2	6	6
				3		1
cox1	13.37	6	6			
				1	3	3
				2	2	2
				3	1	1
cox2	8.08	14	16			

				1	4	5
				2	10	10
				3		1
cox3	11.08	13	15			
				1	3	3
				2	10	11
				3		1
*eleind009**	2.06	0	2			
				2		1
				3		1
matR	2.37	13	15			
				1	3	3
				2	10	10
				3		2
mttB	2.31	28	31			
				1	14	15
				2	14	15
				3		1
nad1	4.06	2	21			
				1		10
				2	2	8
				3		3
nad2	3.94	26	33			
				1	8	10
				2	18	19
				3		4
nad3	2.22	0	14			
				1		5
				2		9
nad4	4.09	19	22			
				1	5	7
				2	14	14
				3		1
nad4L	2.39	9	9			
				1	1	1
				2	8	8
nad5	5.34	10	13			
				1	2	2
				2	8	9
				3		2
nad6	5.29	11	12			
				1	3	4
				2	7	7

				3	1	1
nad7	5.01	25	26			
				1	7	7
				2	18	18
				3		1
nad9	5.58	15	15			
				1	4	4
				2	8	8
				1	3	3
rpl16	8.55	6	11			
				2	5	6
				3	1	5
rps12	3.97	6	6			
				1	2	2
				2	4	4
rps1	366.55	4	5			
				2	3	4
				3	1	1
rps13	0.89	5	5			
				1	2	2
				2	3	3
rps19	2.37	4	6			
				1	1	1
				2	3	3
				3		2
rps2	2.71	8	9			
				1	3	3
				2	5	5
				3		1
rps3	3.78	11	14			
				1	4	5
				2	7	8
				3		1
rps4	1.13	13	15			
				1	3	3
				2	10	10
				3		2
rps7	1.84	1	3			
				1		1
				2	1	1
				3		1
total codon 1					147	175
total codon 2					274	305

total codon 3	10	50
total	431	530

Table 5 Number of differences among mitochondrial components of subunits within chloroids

Subunit	Total sites	Mean no. diff.	Mean no. diff. per site
Nad	8212	1.25	0.000153
Mtt	752	1.33	0.001773
Rpl	557	1.33	0.002394
Cob	1163	2.67	0.002293
Cox	3123	3.66	0.001173
Mat	2036	4.67	0.002292
Ccm	4409	4.69	0.001065
Rps	4902	7.01	0.001431
Atp	3375	11.94	0.003537

Table 6 Comparison of dN dS trees produced for *Eleusine indica*, *Oropetium thomaeum*, *Sporobolus michauxianus*, and *Ananas comosus*

Gene	Total dS tree len.	Total dN tree length
atp1	0.1626	0.0178
atp6	0.0821	0.0292
atp9	0.4828	0.0352
atp4	0.0922	0.0498
atp8	0.2145	0.1357

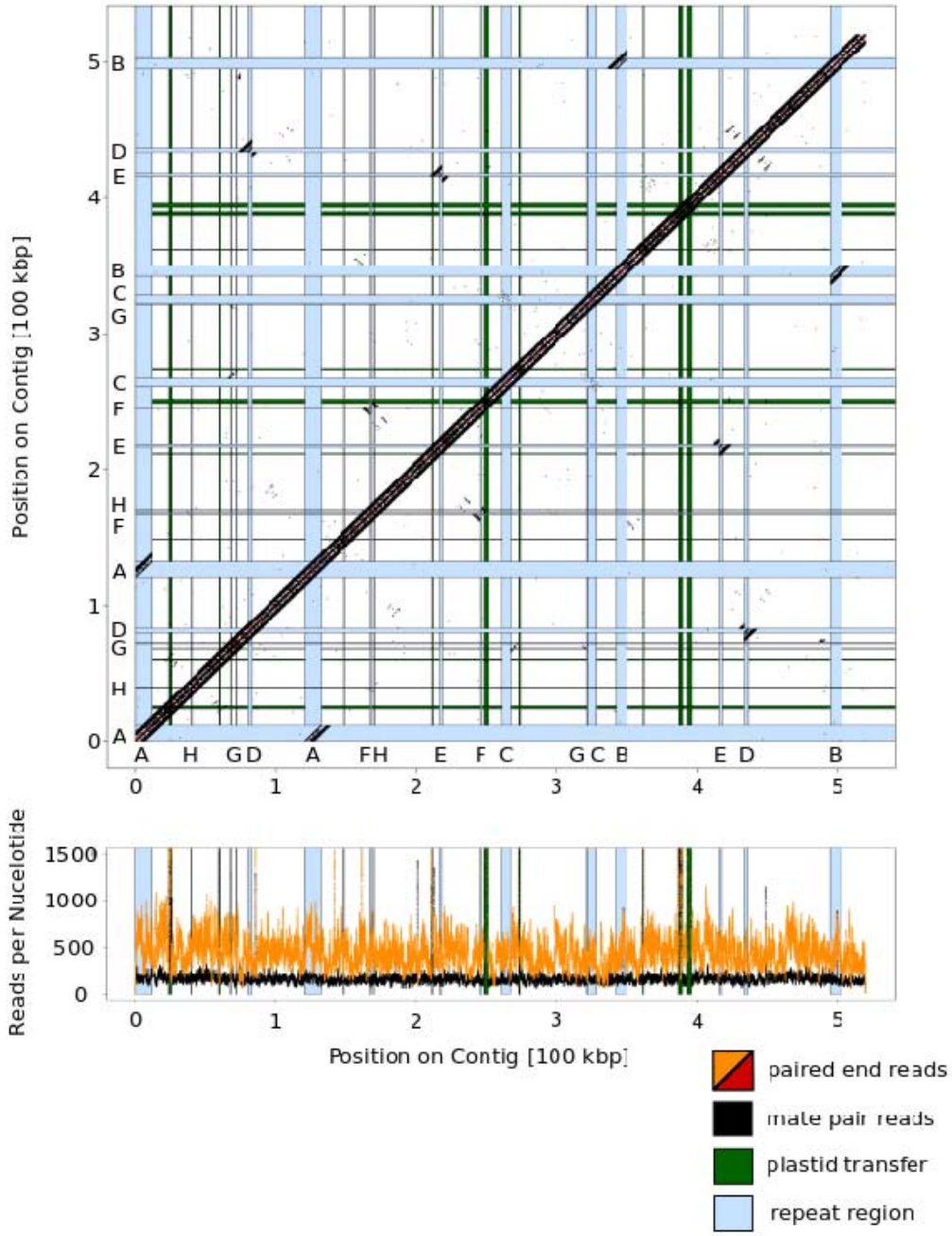


Figure 1 a) Depth plot of mt genome at each nucleotide position (max 8000). b) Connection plot of mate pair and paired end reads on the *Eleusine indica* mt genome. Mate pair information indicates repeat regions are traversed by mate pair reads.

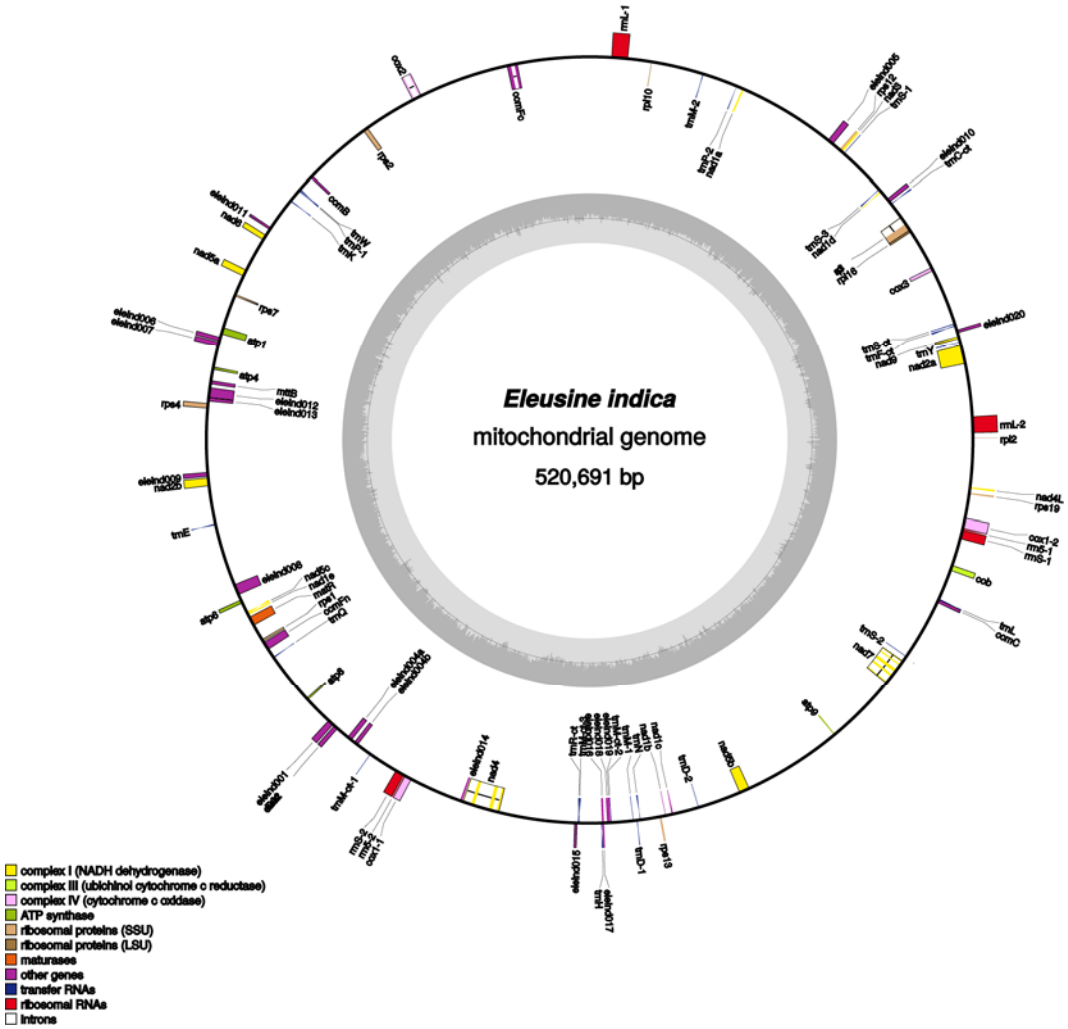


Figure 2 Map of circularized *Eleusine indica* mt genome. Genes marked on the outside of the circle are transcribed counterclockwise, and genes marked on the inside of the circle are transcribed clockwise. Colors are assigned based on function and hypothetical proteins begin with eleind. The inner circle shows GC and AT content; dark gray represents GC and light gray represents AT content

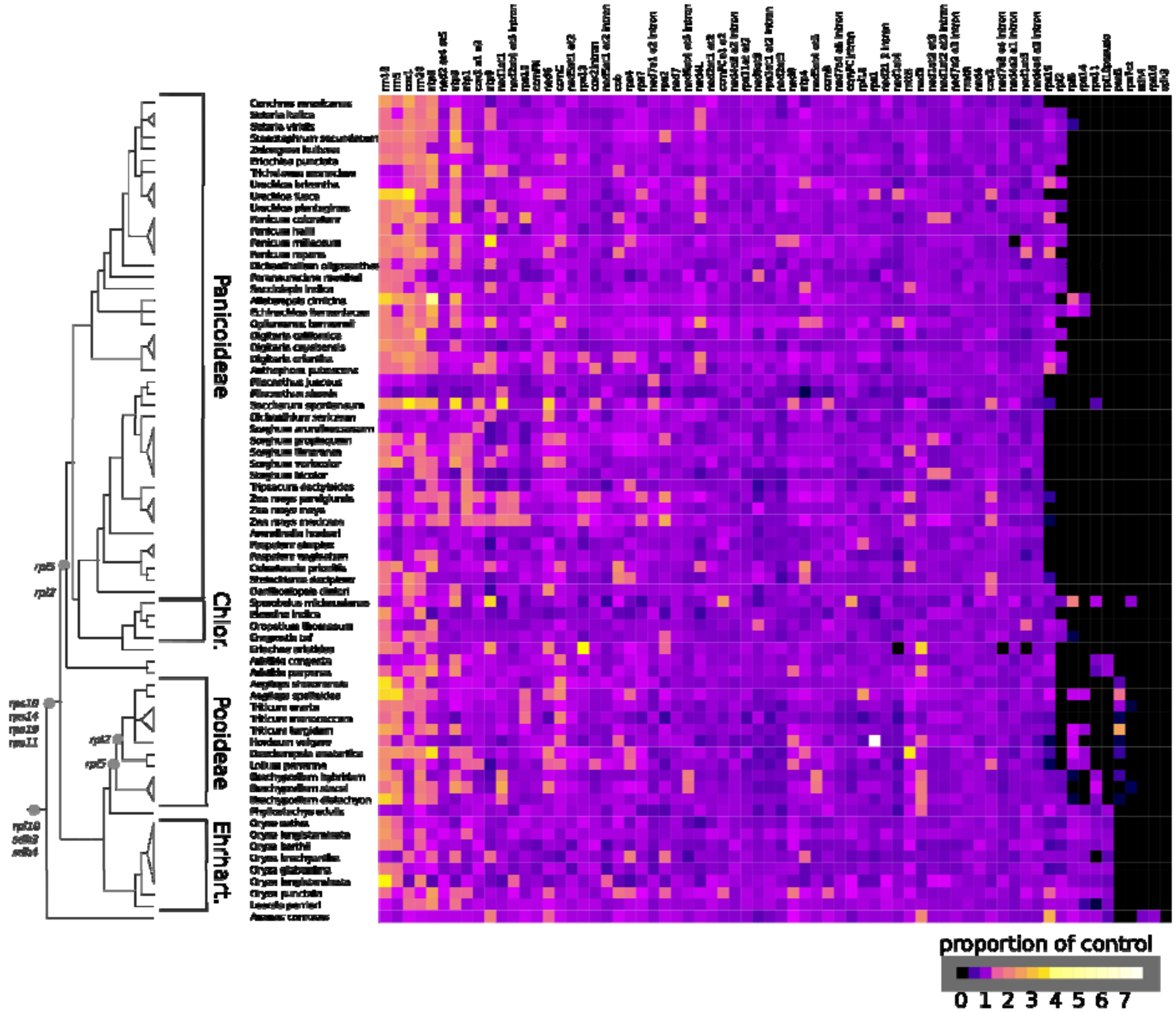
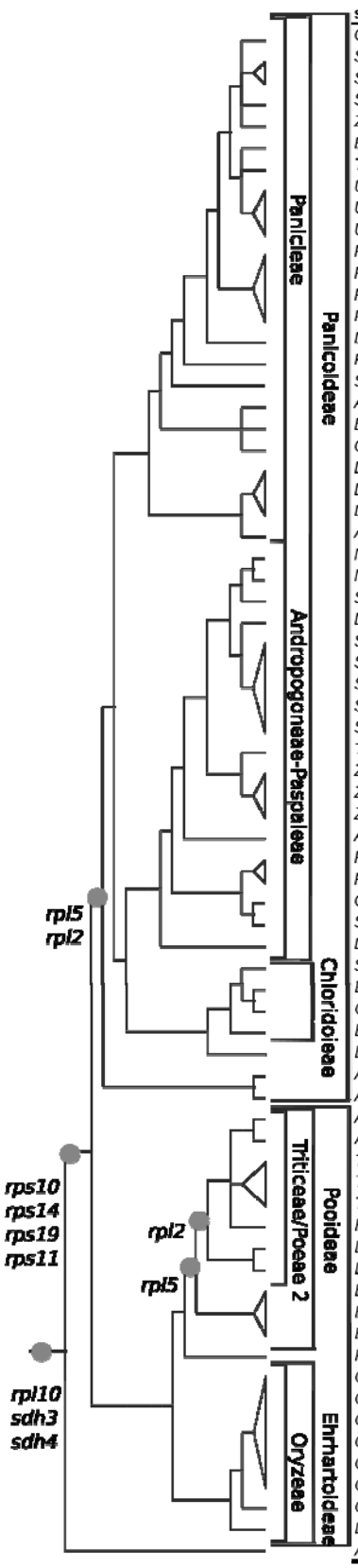

Panicoideae
Chlor.
Pooideae
Ehrhart.
proportion of control
0 1 2 3 4 5 6 7

Figure 3 Heat map of mt gene fragments showing the variable presence of gene sequences in mt genomes within Poaceae and *Ananas comosus*. Taxonomic order was taken from GPWG (2011). Gray dots show functional transfers of genes to the nucleus. The gene *rpl2* is shown in a position deeper than previously reported in the PACMAD clade by Subramanian and Bonen (2006) based on the uniform absence of an intact orf (Table 2, Supplementary Fig.4). Gene transfers taken from the literature are s*dh3, sdh4* (Adams et al. 2001b), *rpl10* (Mower and Bonen 2009; Kubo and Arimura 2010), *rps10* (Adams et al. 2002)*, rps11* (Bergthorsson et al. 2004)*, rps14* (Sandoval et al. 2004; Ong and Palmer 2006), *rps19* (Fallahi et al. 2005; Atluri et al. 2015), *rpl2* (Subramanian and Bonen 2006), and *rpl5* (Sandoval et al. 2004; Wu et al. 2017)



species	avg. depth control	rpl2	rpl5	rpl10	rps11	rps14	rps19	sra
Cenchrus americanus	19.68	?	-	-	-	-	-	SRR2159431
Setaria italica	17.92	○	?	?	?	-	+	SRR2163018
Setaria viridis	15.22	○	○	?	?	-	+	SRR2163019
Stenotaphrum secundatum	14.18	○	?	?	?	?	●	SRR2163117
Zuloagaea bulbosa	17.35	○	-	-	-	-	+	SRR2163119
Eriochloa punctata	14.01	○	?	?	?	?	●	SRR2163123
Tricholaena monachne	20.96	-	-	-	-	-	+	SRR2163183
Urochloa brizantha	20.15	○	-	-	-	-	+	SRR2163211
Urochloa fusca	8.34	?	?	?	?	?	?	SRR2163229
Urochloa plantaginea	9.23	○	?	?	?	?	●	SRR2163231
Panicum coloratum	24.38	-	-	-	-	-	+	SRR2163288
Panicum hallii	21.32	○	-	-	-	-	●	SRR2162783
Panicum miliaceum	7.98	○	?	?	?	?	+	SRR2163427
Panicum repens	15.96	?	-	?	?	-	+	SRR2163429
Dichanthelium oligosanthes	43.24	○	-	-	-	-	+	SRR2162784
Paraneurachne muelleri	11.64	○	?	?	?	?	●	SRR2163452
Sacciolepis indica	18.78	○	-	?	-	-	+	SRR2163432
Alloteropsis cimicina	14.05	?	●	?	?	●	+	SRR2163548
Echinochloa frumentacea	25.27	○	+	-	-	●	+	SRR2162759
Oplismenus burmannii	10.50	?	?	?	?	?	●	SRR2163550
Digitaria californica	29.24	○	-	-	-	-	●	SRR2162855
Digitaria cuyabensis	23.89	-	-	-	-	-	+	SRR2163559
Digitaria eriantha	13.01	○	?	?	?	?	+	SRR2163561
Anthephora pubescens	10.01	○	?	?	?	?	●	SRR2163557
Miscanthus junceus	159.29	-	-	-	-	-	-	SRR396849
Miscanthus sinesis	115.25	-	-	-	-	-	-	SRR1298773
Saccharum spontaneum	34.02	-	-	-	○	-	●	SRR486146
Dichanthium sericeum	30.60	-	-	-	-	-	-	SRR2163562
Sorghum arundinaceaeum	156.88	-	-	-	-	-	-	SRR998981
Sorghum propinquum	325.06	-	-	-	-	-	-	SRR999027
Sorghum timorense	389.03	-	-	-	-	-	-	SRR424217
Sorghum veriscolor	204.46	-	-	-	-	-	-	SRR427175
Sorghum bicolor	249.84	-	-	-	-	-	-	SRR998963
Tripsacum dactyloides	65.98	-	-	-	-	-	-	SRR447807
Zea mays parviglumis	35.96	-	-	-	-	-	?	SRR447755
Zea mays mays	25.31	-	-	-	-	-	-	SRR447927
Zea mays mexicana	30.10	-	-	-	-	-	-	SRR447933
Arundinella hookeri	32.37	-	-	-	-	-	-	SRR2163560
Paspalum simplex	25.18	-	-	-	-	-	-	SRR2162764
Paspalum vaginatum	27.47	-	-	-	-	-	-	SRR2163016
Coleataenia prionitis	12.68	?	?	?	?	?	?	SRR2163565
Steinchisma decipiens	19.42	-	-	-	-	-	○	SRR2163017
Danthoniopsis dinteri	45.23	-	-	-	-	-	+	SRR2163566
Sporobolus michauxianus	162.14	○	●	-	○	-	+	SRR486071
Eleusine indica	568.54	-	-	-	-	-	+	
Oropetium thomaeum	106.32	○	-	-	-	-	+	SRR2083764
Eragrostis tef	310.44	○	-	-	-	-	●	SRR1463402
Eriachne aristidea	6.33	?	?	?	?	?	●	SRR2163567
Aristida congesta	24.56	-	-	●	●	-	+	SRR2163568
Aristida purpurea	9.31	?	?	●	●	?	●	SRR2163569
Aegilops sharonensis	28.71	-	-	-	-	-	○	ERR426382
Aegilops speltoides	272.98	-	+	-	-	●	○	ERR424847
Triticum urartu	248.39	-	-	-	-	-	○	ERR424867
Triticum monococcum	136.90	-	-	-	-	-	○	SRR445609
Triticum turgidum	218.66	-	+	-	-	●	○	ERR463920
Hordeum vulgare	26.11	-	+	-	-	-	?	SRR491411
Deschampsia anatartica	178.63	-	+	-	-	●	○	SRR1158316
Lolium perrenne	69.13	-	+	-	-	●	-	ERR1333447
Brachypodium hybridum	607.80	+	-	-	●	-	+	SRR1800492
Brachypodium stacei	33.66	●	-	-	●	-	+	SRR1801598
Brachypodium distachyon	135.83	●	-	-	●	-	+	SRR191818
Phyllostachys edulis	51.29	●	●	●	●	●	●	ERR105060
Oryza sativa	133.71	+	+	●	●	●	+	SRR3098366
Oryza longistaminata	511.72	+	+	●	●	●	+	SRR2170947
Oryza barthii	12.54	●	●	●	●	●	●	SRR1712907
Oryza brachyantha	28.05	●	●	●	-	●	+	SRR350709
Oryza glaberrima	13.39	●	●	●	●	●	+	SRR1712656
Oryza longistaminata	185.49	+	+	●	●	●	+	SRR1264538
Oryza puctata	19.33	●	+	●	●	●	+	SRR1264539
Leersia perrieeri	1455.81	●	+	○	-	○	+	SRR1528439
Ananas comosus	1328.00	●	+	○	●	●	+	DRR022930

Figure 4 Table of high confidence sequence presence and absence. Presence is based on 3 tiers of evidence. First tier "○" is assigned to genes that have a heat map value of greater than 0.25. Second tier "●" is assigned to genes with continuous coverage across the length of the gene with value greater than ca. -2.0 (Supplementary Fig. 4). Third tier "+" is assigned to genes with sufficient depth and coverage to assemble an intact reading frame. "-" Absence is based on low or no coverage of genes for which the average control is greater than 15. "?" indicates ambiguity related to low depth of coverage for the mitochondrial genome. Grass phylogeny and functional transfers are based on literature cited in Figure 3.

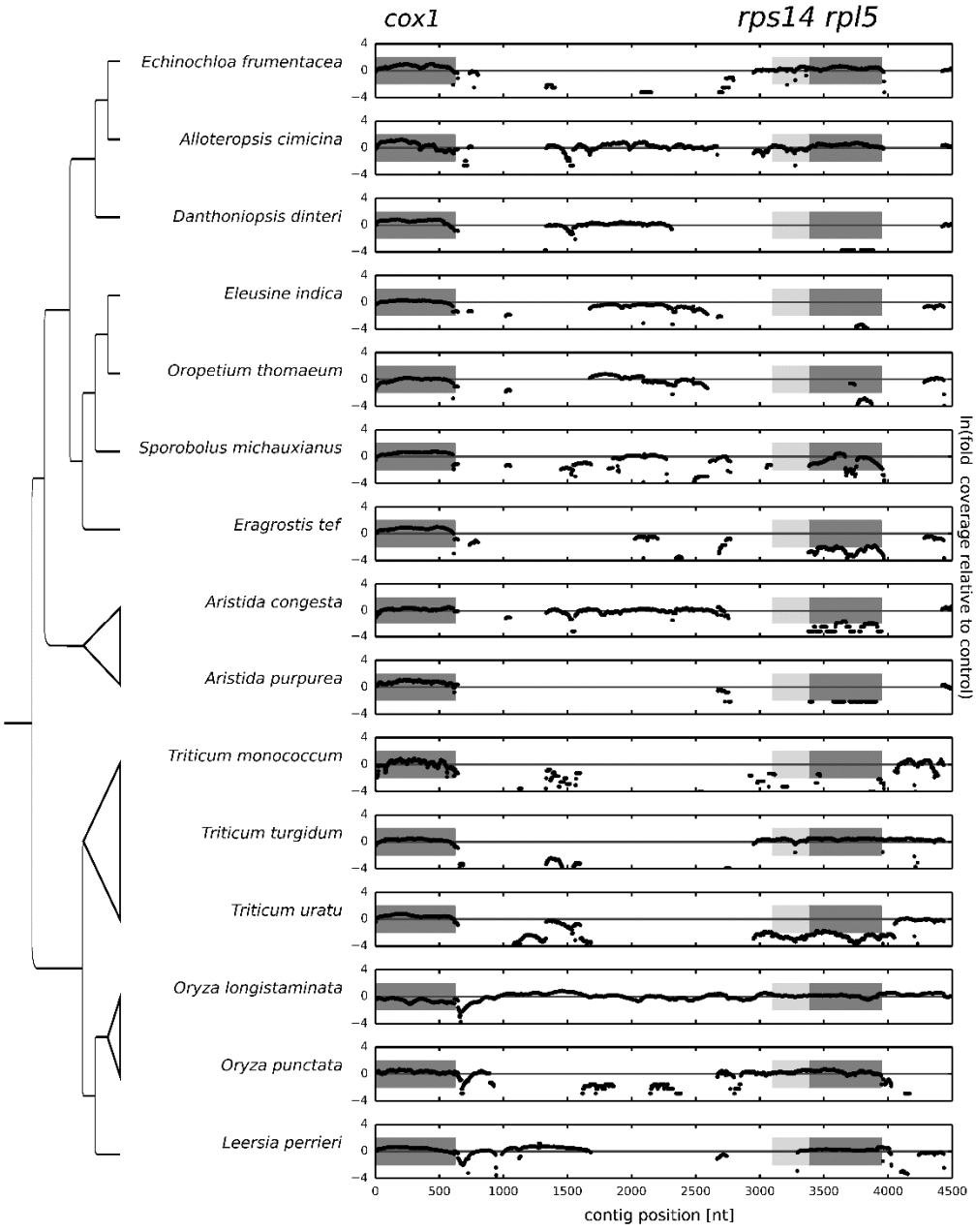
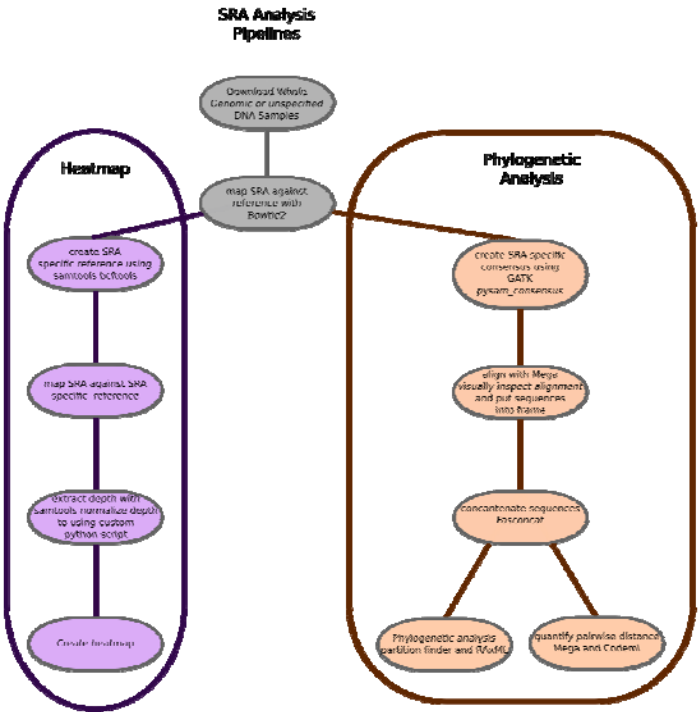
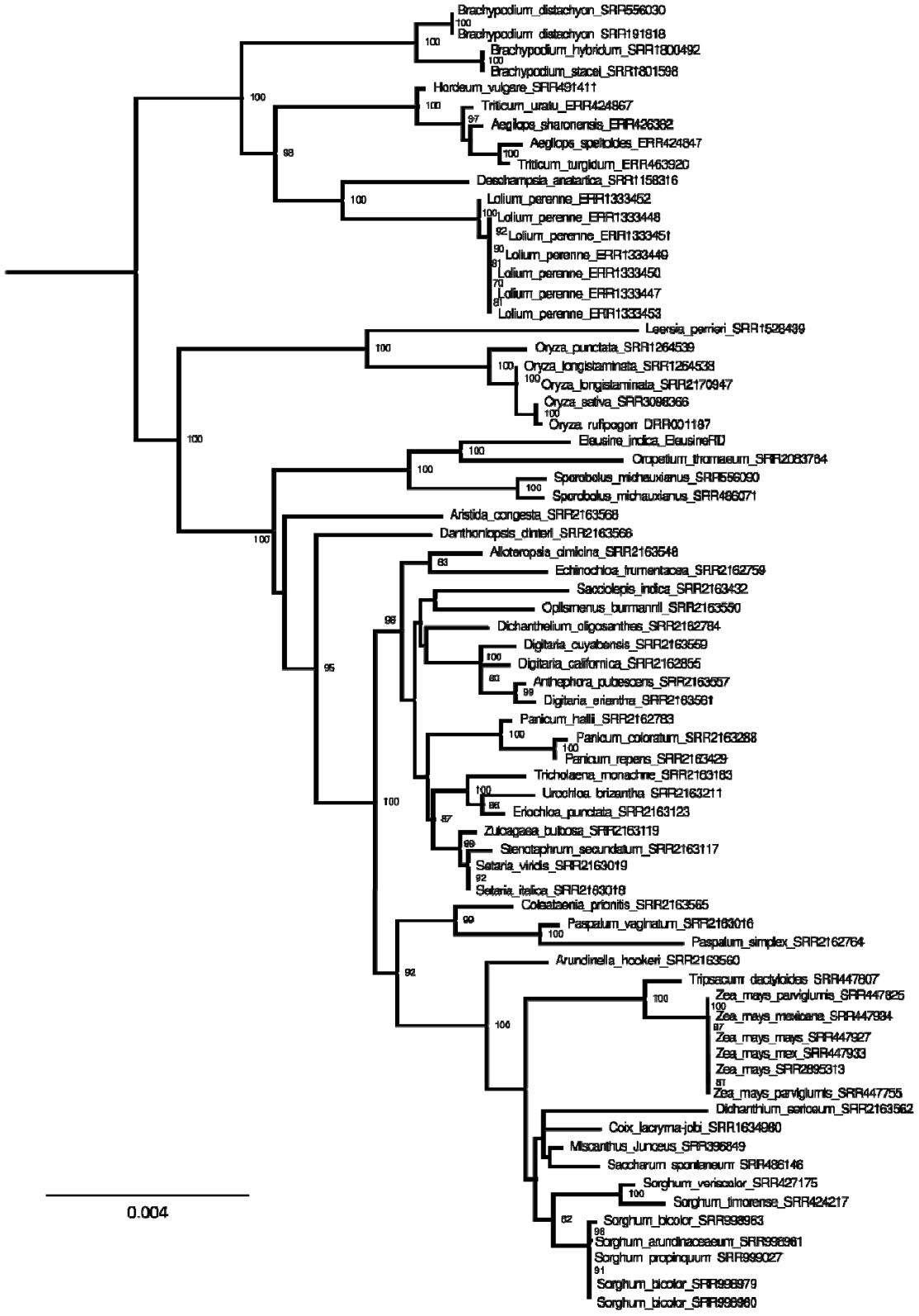

cox1
rps14 rpl5
Echinochloa frumentacea
Alloteropsis cimicina
Danthoniopsis dinteri
Eleusine indica
Oropetium thomaeum
Sporobolus michauxianus
Eragrostis tef
Aristida congesta
Aristida purpurea
Triticum monococcum
Triticum turgidum
Triticum uratu
Oryza longistaminata
Oryza punctata
Leersia perrieri
4
0
−4
0
500
1000
1500
2000
2500
3000
3500
4000
4500
contig position [nt]
ln(fold coverage relative to control)

Figure 5 Plot of the natural log of the mt normalized read depth plus 1e-10. The *cox1* gene fragment serves as the positive control and *rps14* and *rpl5* are variable. The gray boxes mark the borders of the gene on the x-axis for the sequence used as reference *Oryza sativa* (GenBank accession *NC_007886.1:340483-350527*)*;* the y-axis boundaries represent the natural log of a 6-fold deviation from the expected mt coverage. This is helpful in distinguishing the difference between mt and nuclear sequences. For example, the *Triticum uratu* dataset is deeply sequenced, so it is possible to pick up signatures of nuclear transfers, and these must be distinguished from the mt sequence

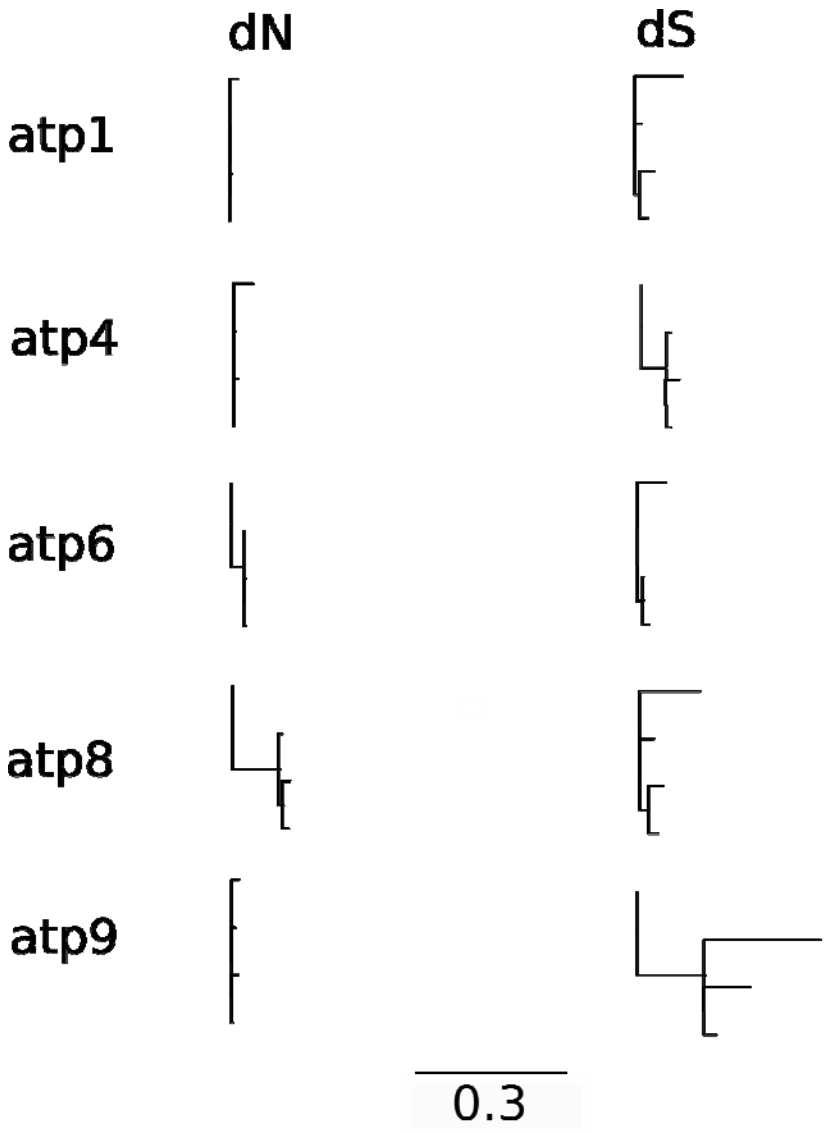


Supplementary Figure 1 Flowchart for handling SRA data

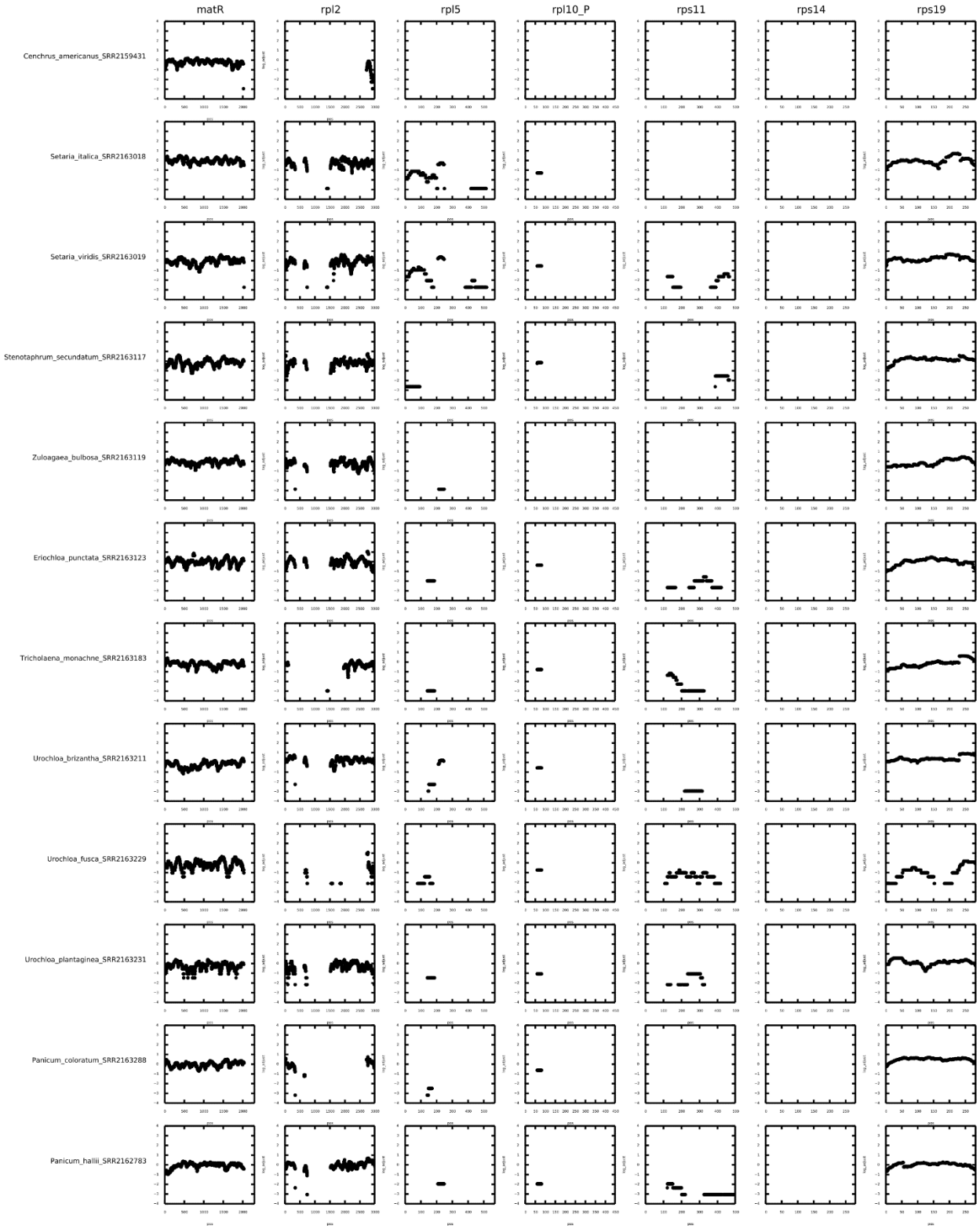

Brachypodium_distachyon_SRR556030
Brachypodium_distachyon_SRR191818
Brachypodium_hybridum_SRR1800492
Brachypodium_stacei_SRR1801598
Hordeum_vulgare_SRR491411
Triticum_urartu_ERR424867
Aegilops_sharonensis_ERR426382
Aegilops_speltoides_ERR424847
Triticum_turgidum_ERR463920
Deschampsia_antarctica_SRR1158316
Lolium_perenne_ERR1333452
Lolium_perenne_ERR1333448
Lolium_perenne_ERR1333451
Lolium_perenne_ERR1333449
Lolium_perenne_ERR1333450
Lolium_perenne_ERR1333447
Lolium_perenne_ERR1333453
Leersia_perrieri_SRR1528499
Oryza_punctata_SRR1264539
Oryza_longistaminata_SRR1264538
Oryza_longistaminata_SRR2170947
Oryza_sativa_SRR3098366
Oryza_rufipogon_DRR001187
Eleusine_indica_EleusineRD
Oropetium_thomaeum_SRR2083784
Sporobolus_michauxianus_SRR556090
Sporobolus_michauxianus_SRR486071
Aristida_congesta_SRR2163568
Danthoniopsis_dinteri_SRR2163566
Alloteropsis_cimicina_SRR2163548
Echinochloa_frumentacea_SRR2162759
Sacciolepis_indica_SRR2163432
Oplismenus_burmannii_SRR2163550
Dichanthelium_oligosanthes_SRR2162784
Digitaria_cuyabensis_SRR2163559
Digitaria_californica_SRR2162855
Anthephora_pubescens_SRR2163557
Digitaria_eriantha_SRR2163561
Panicum_hallii_SRR2162783
Panicum_coloratum_SRR2163288
Panicum_repens_SRR2163429
Tricholaena_monachne_SRR2163183
Urochloa_brizantha_SRR2163211
Eriochloa_punctata_SRR2163123
Zuloagaea_bulbosa_SRR2163119
Stenotaphrum_secundatum_SRR2163117
Setaria_viridis_SRR2163019
Setaria_italica_SRR2163018
Coleataenia_prionitis_SRR2163565
Paspalum_vaginatum_SRR2163016
Paspalum_simplex_SRR2162764
Arundinella_hookeri_SRR2163560
Tripsacum_dactyloides_SRR447807
Zea_mays_parviglumis_SRR447825
Zea_mays_mexicana_SRR447984
Zea_mays_mays_SRR447927
Zea_mays_mex_SRR447933
Zea_mays_SRR2895313
Zea_mays_parviglumis_SRR447755
Dichanthium_sericeum_SRR2163562
Coix_lacryma-jobi_SRR1634980
Miscanthus_Junceus_SRR396849
Saccharum_spontaneum_SRR486146
Sorghum_versicolor_SRR427175
Sorghum_timorense_SRR424217
Sorghum_bicolor_SRR998963
Sorghum_arundinaceaum_SRR998961
Sorghum_propinquum_SRR999027
Sorghum_bicolor_SRR998979
Sorghum_bicolor_SRR998980
0.004

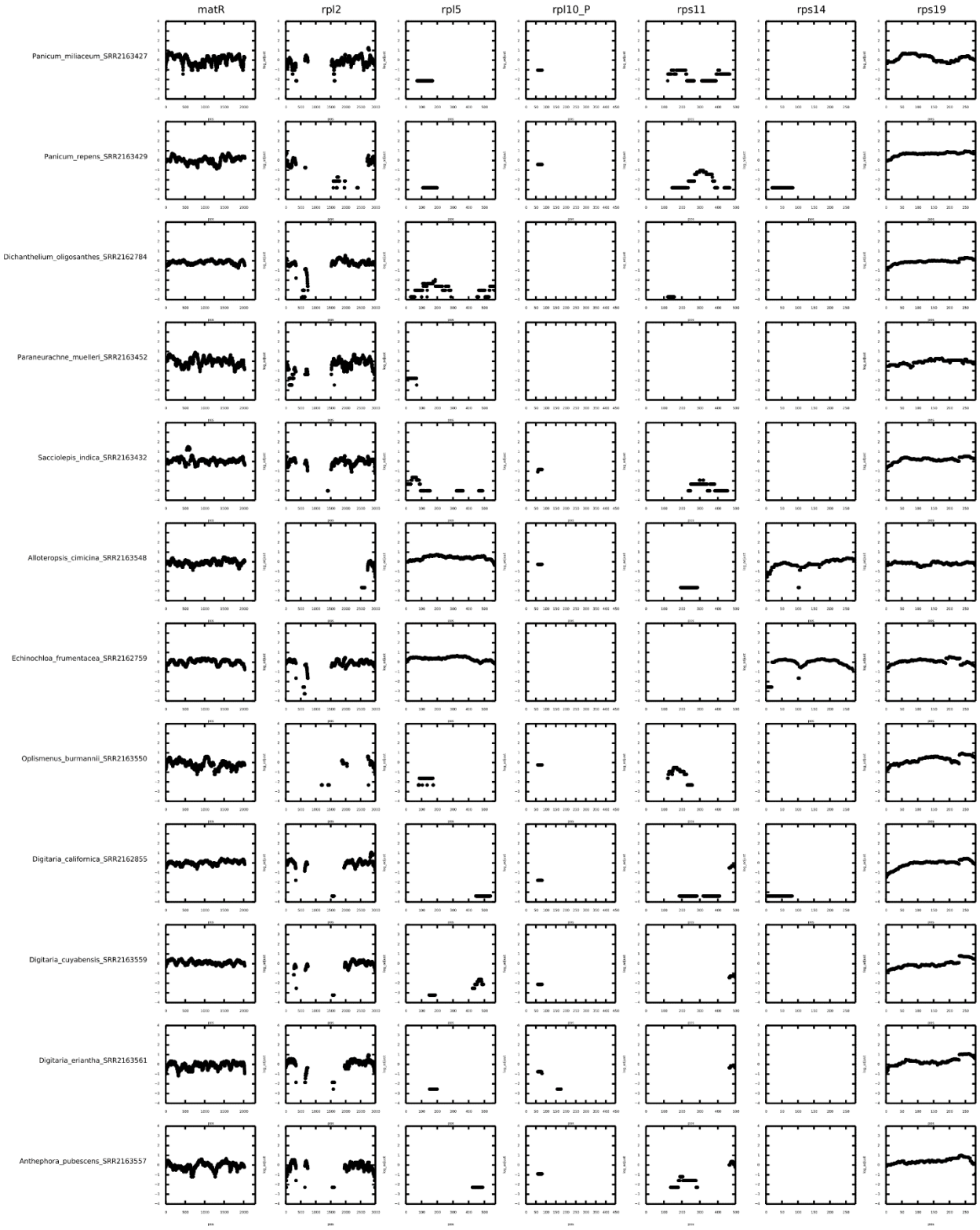
Supplementary Figure 2 Maximum likelihood tree of consensus mt sequences constructed from supermatrix, using codon by gene partitioning scheme. Partitions produced by Partition-Finder and RAxML were run using GTRGAMMA for 100 bootstraps. All sequences had at least 10,000 bp of sequence.

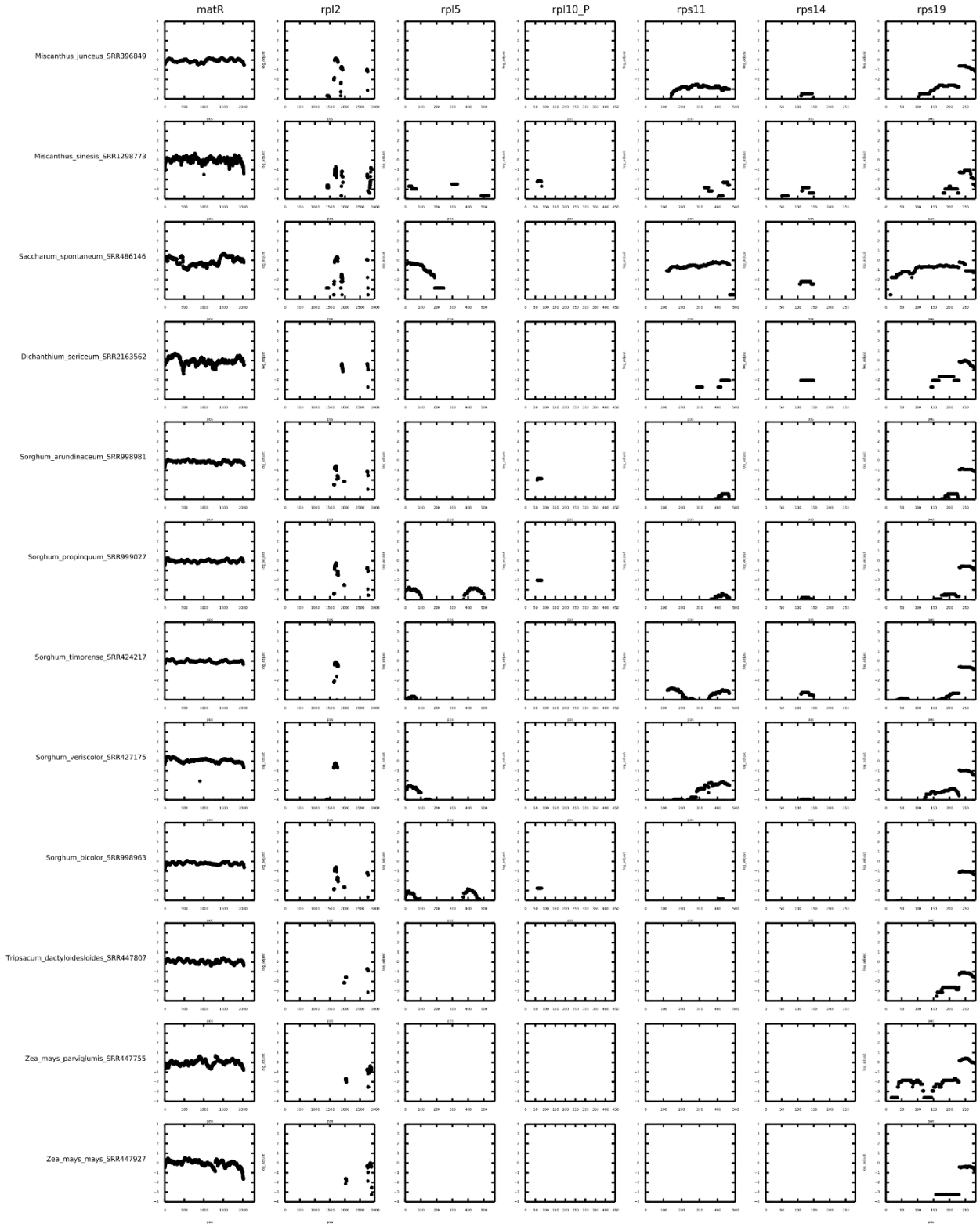
ML tree created from alignment of *atp1*, *atp4*, *atp6*, *atp8*, *atp9*, *ccmB*, *ccmC*, *ccmFC_e1*, *ccmFC_e2*, *ccmFN*, *cob*, *cox1*, *cox2_e1*, *cox2_e2*, *cox3*, *matR*, *mttB*, *nad1_et1*, *nad1_et3*, *nad1_et5*, *nad2_et1*, *nad2_et2*, *nad2_et3*, *nad2_et4*, *nad2_et5*, *nad3*, *nad4L*, *nad4_e1*, *nad4_e2*, *nad4_e3*, *nad4_e4*, *nad5_e3*, *nad5_et1*, *nad5_et2*, *nad5_et4*, *nad5_et5*, *nad6*, *nad7_e1*, *nad7_e3*, *nad7_e4*, *nad7_e5*, *nad9*, *rpl16*, *rps12*, *rps13*, *rps19*, *rps1*, *rps3_e1*, *rps4*, and *rps7*. Bootstrap values of 80 or greater are shown. Tree was rooted using *Ananas comosus* (GenBank accession DRR022930) as an outgroup (not shown here).

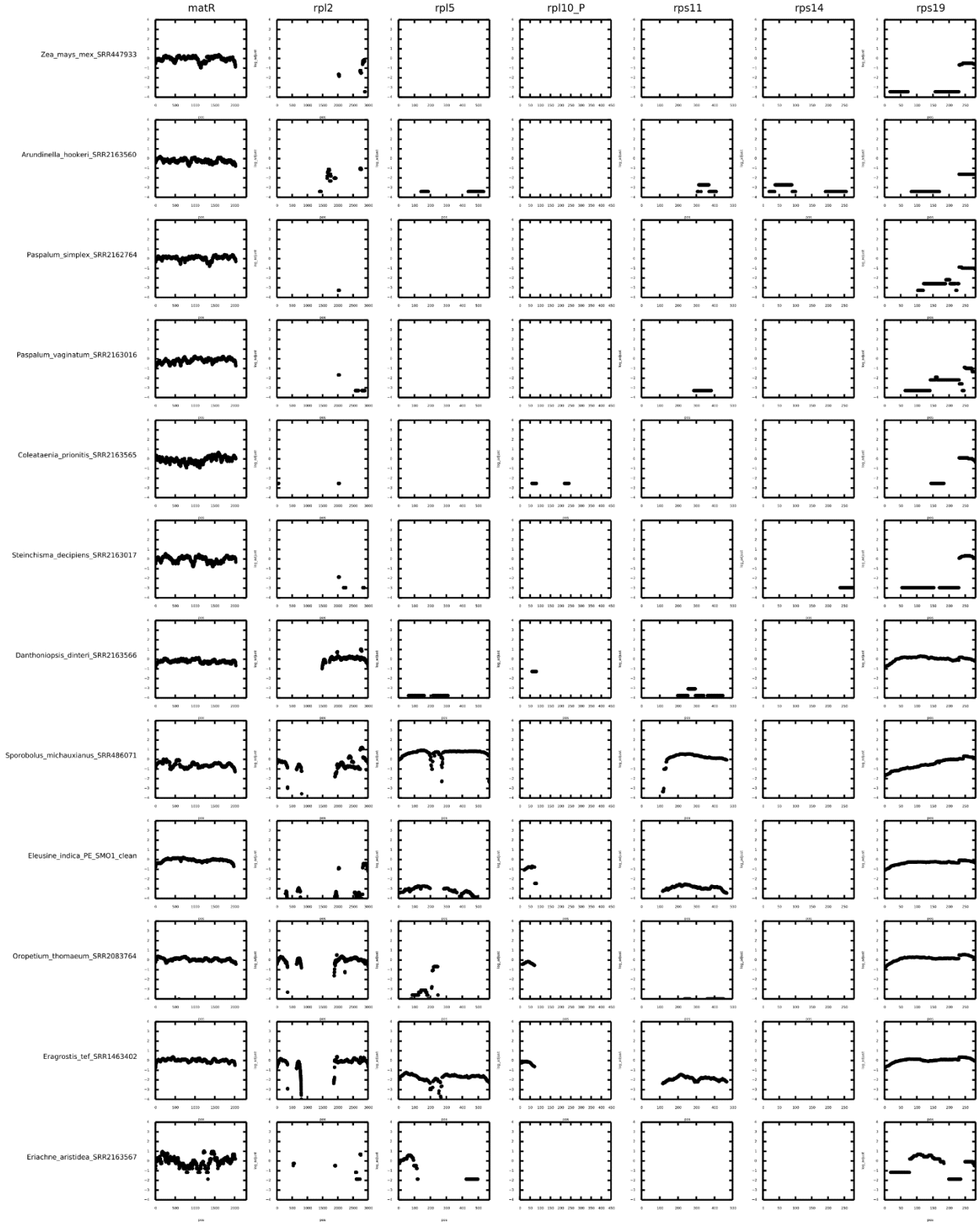


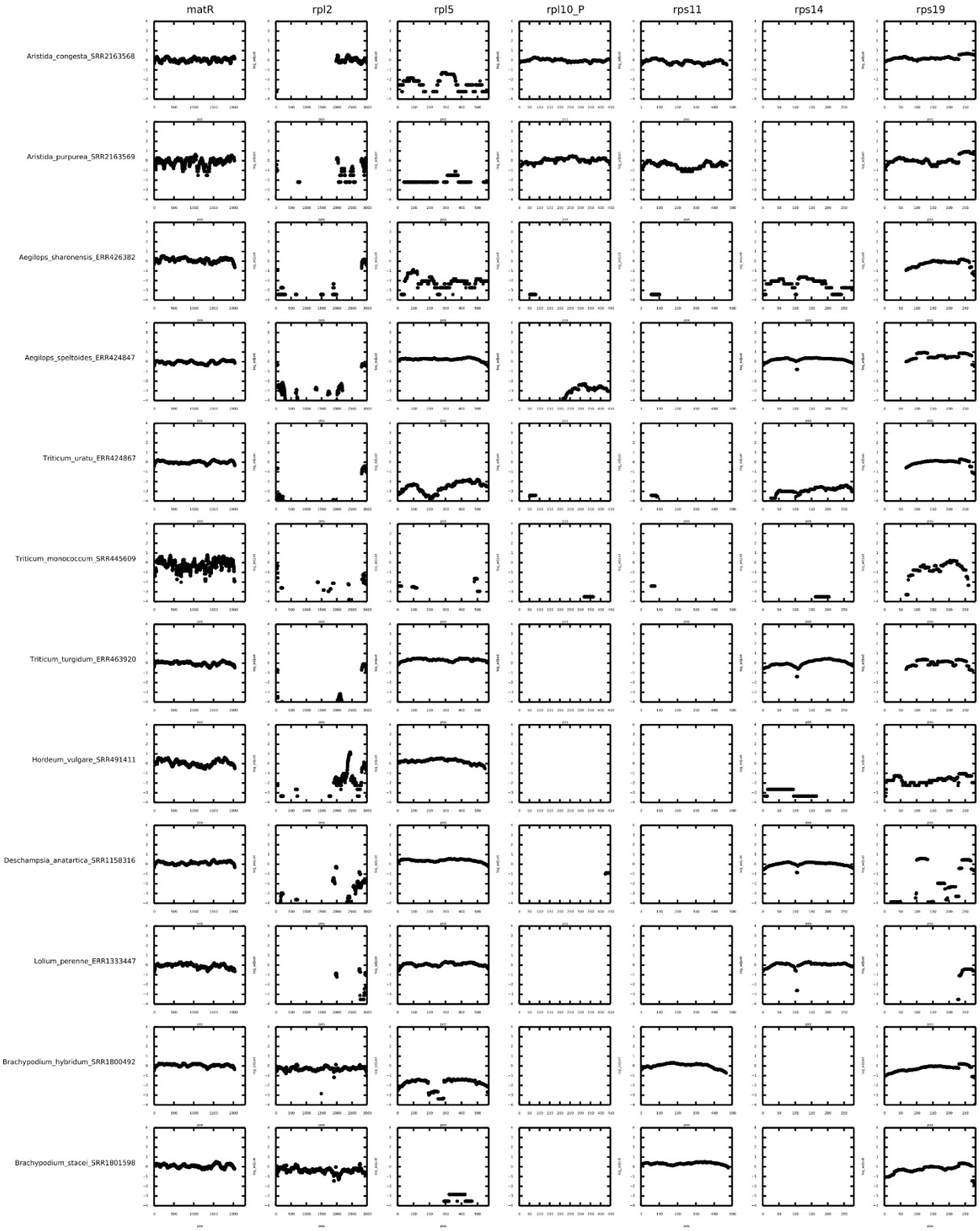
Supplementary Figure 3 Tree lengths taken from codeml to compare the rate of synonymous to non-synonymous substitutions within mt genes that encoded proteins used within atpase subunit

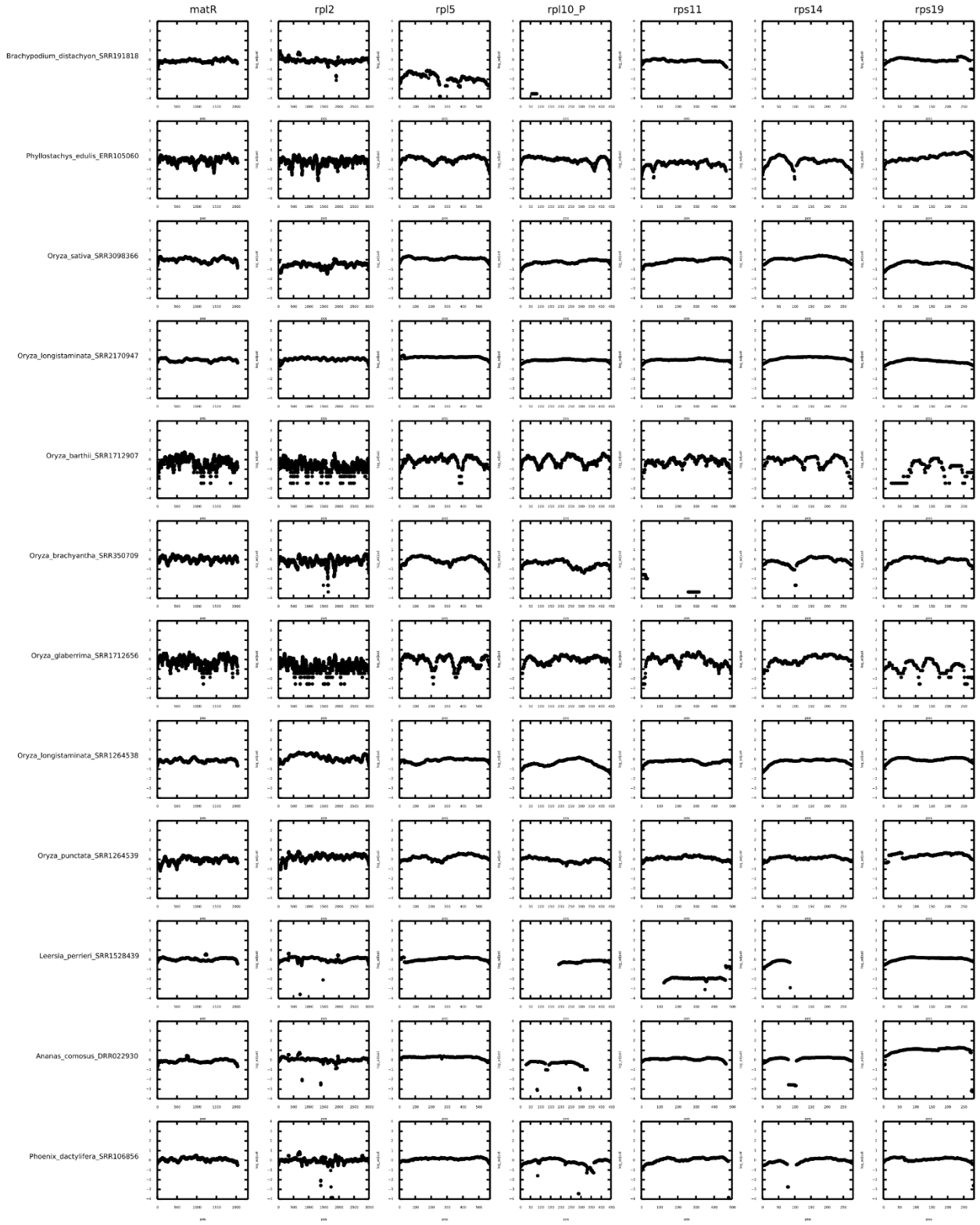

matR
rpl2
rpl5
rpl10_P
rps11
rps14
rps19
Cenchrus_americanus_SRR2159431
Setaria_italica_SRR2163018
Setaria_viridis_SRR2163019
Stenotaphrum_secundatum_SRR2163117
Zuloagaea_bulbosa_SRR2163119
Eriochloa_punctata_SRR2163123
Tricholaena_monachne_SRR2163183
Urochloa_brizantha_SRR2163211
Urochloa_fusca_SRR2163229
Urochloa_plantaginea_SRR2163231
Panicum_coloratum_SRR2163288
Panicum_hallii_SRR2162783


matR
rpl2
rpl5
rpl10_P
rps11
rps14
rps19
Panicum_miliaceum_SRR2163427
Panicum_repens_SRR2163429
Dichanthelium_oligosanthes_SRR2162784
Paraneurachne_muelleri_SRR2163452
Sacciolepis_indica_SRR2163432
Alloteropsis_cimicina_SRR2163548
Echinochloa_frumentacea_SRR2162759
Oplismenus_burmannii_SRR2163550
Digitaria_californica_SRR2162855
Digitaria_cuyabensis_SRR2163559
Digitaria_eriantha_SRR2163561
Anthephora_pubescens_SRR2163557


matR
rpl2
rpl5
rpl10_P
rps11
rps14
rps19
Miscanthus_junceus_SRR396849
Miscanthus_sinesis_SRR1298773
Saccharum_spontaneum_SRR486146
Dichanthium_sericeum_SRR2163562
Sorghum_arundinaceum_SRR998981
Sorghum_propinquum_SRR999027
Sorghum_timorense_SRR424217
Sorghum_veriscolor_SRR427175
Sorghum_bicolor_SRR998963
Tripsacum_dactyloidesloides_SRR447807
Zea_mays_parviglumis_SRR447755
Zea_mays_mays_SRR447927


matR
rpl2
rpl5
rpl10_P
rps11
rps14
rps19
Zea_mays_mex_SRR447933
Arundinella_hookeri_SRR2163560
Paspalum_simplex_SRR2162764
Paspalum_vaginatum_SRR2163016
Coleataenia_prionitis_SRR2163565
Steinchisma_decipiens_SRR2163017
Danthoniopsis_dinteri_SRR2163566
Sporobolus_michauxianus_SRR486071
Eleusine_indica_PE_SMO1_clean
Oropetium_thomaeum_SRR2083764
Eragrostis_tef_SRR1463402
Eriachne_aristidea_SRR2163567


matR
rpl2
rpl5
rpl10_P
rps11
rps14
rps19
Aristida_congesta_SRR2163568
Aristida_purpurea_SRR2163569
Aegilops_sharonensis_ERR426382
Aegilops_speltoides_ERR424847
Triticum_uratu_ERR424867
Triticum_monococcum_SRR445609
Triticum_turgidum_ERR463920
Hordeum_vulgare_SRR491411
Deschampsia_anatartica_SRR1158316
Lolium_perenne_ERR1333447
Brachypodium_hybridum_SRR1800492
Brachypodium_stacei_SRR1801598



Supplementary Figure 4 The natural log of the mt fold depth plus 1e-10 of mapping for each read set across selected genes. The gene matR was used as a positive control

Chapter 3 Sequencing and characterization of the *Del/Tekay* Chromovirus family in *Marshallia obovata* (Asteraceae)

Abstract

Low coverage DNA sequencing data coupled with improved assembly methods are providing new insight into the repetitive fraction of the genomes of non-model organisms. In this study I explore the potential of these new approaches to characterize the *Del/Tekay* elements (chromodomain-containing *Ty3/Gypsy* retrotransposons) of the North American wildflower species *Marshallia obovata*. The *Marshallia Del/Tekay* chromoviruses are 8.3 Kbp in length with direct repeats of 1,675 bp and 14 conserved domains across a 4,506 bp gag-pol region. Read depth comparisons suggest that *Del/Tekay* elements are not highly abundant in *M. obovata* with approximately 144 unique copies per haploid genome of 6.5 Gbp (<0.02 %). Sequence variation was observed at less than 1 % of nucleotide positions overall with 60 % of variants occurring outside of any coding region. These results illustrate the utility of low coverage sequencing data for fine scale analyses of transposable elements present at low to moderate copy numbers within the genomes of non-model organisms. The *Del/Tekay* element sequences described here are the first for the plant family Asteraceae and are the first transposable elements of any kind described for the genus *Marshallia*. These initial results provide for future work on the evolutionary dynamics of *Del/Tekay* elements more broadly in the genus *Marshallia* and Asteraceae.

Introduction

Transposable elements (TEs) are a large and dynamic component of many plant genomes (Bennetzen 2002; Feschotte et al. 2002; Vitte & Panaud 2005; Wei et al. 2009). In essence they are selfish genetic elements that proliferate through copy or cut and paste mechanisms, potentially increasing genome sizes over several orders of magnitude (Bennetzen & Wang 2014; Michael 2014). Their activity in genomes can facilitate adaptive evolution in numerous ways (Kidwell & Lisch 2001; Oliver et al. 2013), but they are also disruptive and trigger silencing countermeasures from the host organism such as epigenetic surveillance or outright elimination (Lee & Kim 2014; Michael 2014). The consequent rise and fall of distinct TE families are notable evolutionary phenomena in their own right and can also inform host phylogeny. Transposable element family abundance, locus-specific insertions and horizontal transfer events can distinguish clades of host species and be used to infer evolutionary relationships (Shedlock et al. 2004; Schaack et al. 2010; Hertweck 2013; Piednoël et al. 2013; Dodsworth et al. 2015) but have been underutilized in general as detailed repetitive element characterizations are uncommon outside of well-studied models or other species with abundant genomic data resources. Recent developments in both sequencing technology and short read assembly and analysis methods are improving our understanding of the diversity and abundance of TEs in non-model organisms and the analyses of these data in an evolutionary framework (Novák et al. 2010; Muñoz-Diez et al. 2012; Hertweck 2013).

In this study I explore the potential of low coverage genomic sequencing to provide detailed characterization of a distinct clade of large (>8 kbp), low-abundance transposable elements within a non-model organism. I leverage paired-read information in assembly and mapping to accurately determine long terminal repeat (LTR) and flanking region sequences,

clearly circumscribe the TE lineage, and estimate its variability within the host genome. Structural annotation and phylogenetic analyses of the validated consensus sequence provide a starting point for additional molecular evolutionary studies focused on the element group in other individuals and related species.

Materials and Methods

Study System

This study focused on *Marshallia obovata* (Walt.) Beadle & F.E. Boynton (Asteraceae), a perennial wildflower native to the southeastern United States. *Marshallia obovata* is diploid (2n = 2x = 18) with a genome size of approximately 6.6 Gb (1 C = 6.79 pg) determined by flow cell cytometry (personal communication, T. Garnatje, Institut Botànic de Barcelona). DNA was extracted from fresh leaf material of a wild-collected plant from Macon County, Alabama (Hansen 4956, AUA) using the modified CTAB protocol of Doyle & Doyle (1987). Samples were submitted to the Genomic Services Lab, Hudson Alpha Institute for Biotechnology (Huntsville, AL), where paired-end libraries with a mean insert size of 281 bp were prepared and sequencing of 39,124,284 100 bp reads was performed on an Illumina Hi-Seq 2000 platform.

Assembly, Validation, and Annotation

Contig assembly was performed with Ray (Boisvert et al. 2010, v2.3.1) using kmer length 31 and minimum 30 bp reads filtered using Sickle (Joshi & Fass 2011, v1.210) to exclude

bases with a Sanger quality score less than 30. Putative TEs containing contigs were identified using RepeatMasker (Smit et al. 2010, v3.2.7) and iteratively extended in the assembler PriceTI (Ruby et al. 2013, v1.0.1) with stringent (99 %) matching. The LTRharvest (Ellinghaus et al. 2008) and LTRdigest (Steinbiss et al. 2009) modules of Genome Tools (Gremme et al. 2013, v1.5.3) were used to identify intact Class I retrotransposons with long terminal repeats (LTRs) and complete gag-pol regions. A candidate transposon sequence was selected for further analysis based on its large size combined with moderate read coverage to explore the upper limits of repetitive element assembly and annotation in this genome skimming dataset.

The LTR regions of the partial transposon contig were extended with local assembly of read pairs until the 50 sequence of LTR1 matched the gag-pol-LTR2 boundary and the 3′ end of LTR2 matched the LTR1-gag-pol boundary. The original genomic reads were then re-mapped to a consensus sequence in Bowtie 2 (Langmead & Salzberg 2012, v.2.2.4), allowing only correctly paired reads and end-to-end alignment enforced to validate the assembly overall, determine a consensus sequence, and extract variant calls. Annotation of element components was accomplished with Genome Tools modules LTRdigest and sketch, and sequence comparisons were performed with BLAST+ (NCBI Resource Coordinators 2014) and Artemis (Rutherford et al. 2000, v16.0.0).

Element Classification

Sequence features of the validated consensus including LTR length and sequence domain presence/order were compared to known TEs and found to be consistent with the chromoviral branch of the *Ty3/Gypsy* LTR retroelement family. Phylogenetic analyses were also conducted

for a fine-scale classification of the *Marshallia* retroelements. Representative sequences were obtained from datasets referencing chromoviruses (Neumann et al. 2011; Llorens et al. 2011; Kolano et al. 2013; Domingues et al. 2012) and from sequence similarity searches of NCBI databases using the *Marshallia* element gag-pol region as a query.

Amino acid sequence alignment for a portion of the reverse transcriptase domain was performed in ClustalW (Thompson et al. 1994, v2.1) with manual adjustments where necessary. Substitution model choice and maximum likelihood analyses including 300 bootstrap replicates were carried out with RAxML (Stamatakis 2014, v8.1.5). Phylogenetic trees were visualized with FigTree (Rambaut 2010, v1.3.1).

Element Circumscription, Variation, and Copy Number

A less stringent alignment of genomic reads, allowing up to 39 nucleotide differences or indels relative to the reference, was created with Bowtie 2 with relaxed mapping (the minimum default score for mapping adjusted with the formula minimum score = -0.6 + -2x (length of read), with forced read pairing and no match bonus) to distinguish reads belonging to this retroelement group from more distantly related TEs. Reads were binned by the observed number of nucleotide or indel differences and plotted in R (R Core Team 2013, v.3.0.2). A high-stringency map (forced read pairing, no match bonus in default smithwaterman alignment) was produced with Bowtie 2 to assess intragenomic variation. The resulting alignment was converted into pileup format with SAMtools (Li et al. 2009, v1.2). A custom R script (courtesy of Eric Archer, NOAA Fisheries) was used to extract nucleotide frequencies at every position, including only those single nucleotide polymorphism (SNP) variants above 5 % and supported by a

minimum of 5 reads. Insertion or deletion (indel) variants coded as CIGAR strings in the SAM file produced by mapping were characterized and counted in Tablet (Milne et al. 2013, v1.14.10.21) using the default feature threshold of 10 CIGAR strings per site.

Average reads per kilobase (RPK) for single copy nuclear genes was calculated by local mapping (no read pairing, match bonus default value, default smith-waterman alignment) genomic reads to 897 unique COSII markers (Wu et al. 2006) identified and extracted from an unpublished *Marshallia obovata* transcriptome. The mean of all COSII RPKs was compared to the RPKs obtained from a high stringency mapping and a local mapping of the genomic reads to the element consensus. Number of copies was estimated by dividing the *Del/Tekay* RPK by the average COSII RPK value.

Results

The fully assembled *Marshallia obovata Del/Tekay* chromovirus consensus sequence (NCBI accession KX396599) is 8,308 bp in length with a 4,506 bp gag-pol region flanked by direct repeats of 1,675 bp (Fig. 1). Annotation yielded 14 conserved regions or domains including a predicted 18 bp primer binding site (PBS) for tRNA (iMet) initiated replication. The gag-pol region is a single, intact open reading frame containing the expected catalytic domains providing for the retrotransposon gag protein, zinc finger, retroviral aspartyl protease, aspartate protease, reverse transcriptase, Ribonuclease H, integrase core domain, and a chromodomain. A polypurine tract and uracil-rich, putative U-box were detected between the end of the gag-pol region and the start of LTR2. Additional features identified included universal minicircle sequence binding protein (UMSBP, 2996–3163), a second zinc knuckle (zf-CCHC, 3116–3167),

and an Arginine methyltransfersase-interacting protein (AIR1, 3065–3175). The length of the terminal repeat falls within reported values (1.1–4.4kbp) for the *Del/Tekay* group of chromoviruses (Llorens et al. 2011), and the linear order of domains further supports *Del/Tekay* assignment (Weber et al. 2013). Phylogenetic analyses of the reverse transcriptase domain also suggest the *Marshallia obovata* elements are members of the *Del/Tekay Ty3/Gypsy* chromoviruses and most closely related to sequences from other asterids (Fig. 2).

Low-stringency mapping identified a total of 14,946 reads that matched the *Del/Tekay* consensus sequence at a minimum of 60 nucleotides out of 100 (Fig. 3). The majority of these reads, 13,567 or 91 %, had six or fewer nucleotide differences from the consensus. Approximately 600 reads, not included in the consensus sequence but matching at the 70–85 % level, were identified and localized to conserved LTR motifs and reverse transcriptase or ribonuclease H domains of other TE families.

The average read depth in the more stringent map produced for variant analysis was 85.4 reads per nucleotide with a maximum coverage of 157 reads per nucleotide. Within this map a total of 50 nucleotide polymorphisms were detected, 29 in non-coding sequence including the LTRs and 21 in the coding regions (Table 1). SNP frequencies ranged from 0.05–0.50 of total reads with roughly half of the SNP frequencies under 0.1 and nearly half above 0.25. Within the coding region, there were 13 polymorphisms in the ORF at large and 8 in the conserved domains. Of the total 11 non-synonymous variants, 8 were identified outside the conserved domains and 1 each in the Chromo, RVP2, and UMSBP domains, with frequencies of 0.062, 0.0496, and 0.440, respectively. No variants that resulted in a stop codon were detected. A total of 21 indel variants were identified, 19 in the non-coding portion of the element, 1 within the ORF at large, and one

within Retrotrans_gag, all consisting of slight variations in the length of mono- or dinucleotide repeats.

In the maps prepared for copy number assessment, the coverage depth across the entire *Del/Tekay* element had a mean of 1,633 reads per kilobase (RPK) under high stringency conditions and 2,667 RPK with local alignment allowing unpaired reads. The average read depth for 897 COSII single-copy nuclear genes (local alignment, unpaired reads allowed) was 18.55 RPK giving a *Del/Tekay* copy number estimate of 2,667 RPK/18.55 RPK = 144 copies.

Discussion

Several studies have used low coverage sequencing and consensus methods to assemble and characterize the transposable element fraction of plant genomes at the level of family and superfamily (Novák et al. 2010; Hertweck 2013; Staton & Burke 2015). Here I demonstrate an extension of this approach that allows for a more detailed assessment of TE composition including full-length LTR sequences and estimates of intragenomic variation and abundance.

The 8.3 kbp repetitive element described here corresponds to a uniform and clearly distinct set of LTR retrotransposons within the *Marshallia obovata* genome. Domain annotation and phylogenetic analyses identify them as members of the *Del/Tekay* lineage of *Ty3/Gypsy* chromoviruses. Although the *Ty3/Gypsy* elements are suggested to be particularly active in Asteraceae (Renaut et al. 2014; Staton & Burke 2015), this is the first specific report of a *Del/Tekay* retrotransposon in the family and the first transposable element of any kind described for the genus *Marshallia*.

Sequence variation for the *Del/Tekay* element family was relatively low, with slightly less than 1 % of alignment positions containing a variant nucleotide above a 0.05 frequency threshold. The majority of those variants (29/50, 59 %) occur in non-coding regions, 10 of the remaining 21 are synonymous substitutions, and 8 of the 11 non-synonymous substitutions occur outside of conserved domains. Nearly half of the SNPs occur at low frequency (<0.1) and would be consistent with more recent proliferation of active elements. *Ty3/Gypsy* elements appear to be actively expanding in Asteraceae genomes (Staton & Burke 2015) and the *Del/Tekay* family in *Marshallia* may be experiencing similar growth. Several presumably older variants occur at a significantly higher frequency with 23 above 0.25 and 9 of those above 0.45. These markers may be useful for tracking diversification events in the *Del/Tekay* element family at the population level in *Marshallia obovata* or further back in time above the species level. Of 21 verified indel variants, 19 occur in non-coding positions and 2 in the gag-pol region. Our variant calling process used a cutoff of 0.05 and is therefore insensitive to any mutations (including missense) at individual loci, but the general pattern of most variation occurring outside the ORF region would not likely change with less stringent variant calls. Additional resolution might also be obtained with a refined variant calling method such as indel realignment.

Del/Tekay elements exhibit a wide range in copy number across plant genomes ranging from 46 distinct insertions in *Arabidopsis thaliana* (Du et al. 2010) to over 10,000 in *Pisum sativum* (Neumann et al. 2011). The comparison of mean COSII to *Del/Tekay* RPK values suggests a copy number in *Marshallia obovata* of around 144 distinct loci, assuming all insertions have been diploidized. If all occurrences are unique on individual chromosomes, the coverage depth suggests 288 distinct retrotranspositions. Given the 13 Gbp diploid genome, this represents less than 0.02 % of the nuclear genome and would be considered a low abundance

retroelement, although copy number may be underestimated here since *Del/Tekay* transposons are known to target regions of heterochromatin (Mlinarec et al. 2016) which are likely to be under-sampled in Illumina sequencing experiments (van Dijk et al. 2014).

The *Marshallia Del/Tekay* assembly illustrates how low coverage genomic sequencing data can be used for finescale analyses of TEs. This could be particularly valuable in non-model organisms that often contain new classes of TE and where *de novo* assembly approaches are preferable to targeted techniques that rely on known element families. The assembly and characterization of the *Marshallia Del/Tekay* element family here is the first of its kind for the Asteraceae, a large and important family of plants, as well as for the genus *Marshallia*, an interesting group of wildflowers because little is known about species relationships. This assembly also illustrates a potentially underutilized aspect of increasingly available genomic data and paves the way for broader and more detailed investigations of retrotransposon evolutionary dynamics.

References

Bennetzen JL (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica 115:29–36

Bennetzen JL, Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annual Review of Plant Biology 65:505–530

Boisvert S, Laviolette F, Corbeil J (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. Journal of Computational Biology 17:1519–1533

van Dijk EL, Jaszczyszyn Y, Thermes C (2014) Library preparation methods for next-generation sequencing: tone down the bias. Experimental Cell Research 322:12–20

Dodsworth S, Chase MW, Kelly LJ, et al. (2015) Genomic repeat abundances contain phylogenetic signal. Systematic Biology 64:112–126

Domingues DS, Cruz GMQ, Metcalfe CJ, et al. (2012) Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. BMC Genomics 13:137

Doyle JJ, Doyle JL (1987) CTAB DNA extraction in plants. Phytochemical Bulletin 19:11–15

Du J, Tian Z, Hans CS, et al. (2010) Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. The Plant Journal 63:584-598

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics 9:18

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics 3:329–341

Gremme G, Steinbiss S, Kurtz S (2013) GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Transactions on Computational Biology and Bioinformatics 10:645–656

Hertweck KL (2013) Assembly and comparative analysis of transposable elements from low coverage genomic sequence data in Asparagales. Genome 56:487–494

Joshi NA, Fass JN (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.210). Available at https://github.com/najoshi/sickle.

Kidwell MG, Lisch DR (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution 55:1–24

Kolano B, Bednara E, Weiss-Schneeweiss H (2013) Isolation and characterization of reverse transcriptase fragments of LTR retrotransposons from the genome of *Chenopodium quinoa* (Amaranthaceae). Plant Cell Reports 32:1575–1588

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–359

Lee SI, Kim NS (2014) Transposable elements and genome size variations in plants. Genomics and Informatics 12:87–97

Li H, Handsaker B, Wysoker A, et al. (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Llorens C, Futami R, Covelli L, et al. (2011) The gypsy database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Research 39:D70–4

Michael TP (2014) Plant genome size variation: bloating and purging DNA. Briefings in Functional Genomics 13:308– 317

Milne I, Stephen G, Bayer M, et al. (2013) Using tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics 14:193–202

Mlinarec J, Franjević D, Harapin J, Besendorfer V (2016) The impact of the *Tekay* chromoviral elements on genome organisation and evolution of *Anemone* s.l. (Ranunculaceae). Plant Biology 18:332–347

Muñoz-Diez C, Vitte C, Ross-Ibarra J, Gaut BS, Tenaillon MI (2012) Using nextgen sequencing to investigate genome size variation and transposable element content. In Plant Transposable Elements (edited by MA Grandbastien, JM Casacuberta), vol. 24 of Topics in Current Genetics pp. 41–58. Springer Berlin Heidelberg.

NCBI Resource Coordinators (2014) Database resources of the national center for biotechnology information. Nucleic Acids Research 42:D7–17

Neumann P, Navrátilová A, Koblížková A, et al. (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mobile DNA 2:1.

Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in nextgeneration sequencing data. BMC Bioinformatics 11:378.

Oliver KR, McComb JA, Greene WK (2013) Transposable elements: powerful contributors to angiosperm evolution and diversity. Genome Biology and Evolution 5:1886–1901

Piednoël M, Carrete-Vega G, Renner SS (2013) Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. The Plant Journal 75: 699–709

R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut A (2010) FigTree. Ver.1.3.1. Institute of Evolutionary Biology, University of Edinburgh.

Renaut S, Rowe HC, Ungerer MC, Rieseberg LH (2014) Genomics of homoploid hybrid speciation: diversity and transcriptional activity of long terminal repeat retrotransposons in hybrid sunflowers. Philosophical transactions of the Royal Society of London. Series B, Biological Sciences 369:20130345.

Ruby JG, Bellare P, Derisi JL (2013) PRICE: software for the targeted assembly of components of (meta) genomic sequence data. G3 3:865–880

Rutherford K, Parkhill J, Crook J, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945

Schaack S, Gilbert C, Feschotte C (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends in Ecology and Evolution, 25:537–546

Shedlock AM, Takahashi K, Okada N (2004) SINEs of speciation: tracking lineages with retroposons. Trends in Ecology Evolution 19:545–553

Smit AFA, Hubley R, Green P (2010) 2010 RepeatMasker open-3.0. URL: http://www.repeatmasker.org.

Stamatakis A (2014) Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Staton SE, Burke JM (2015) Evolutionary transitions in the Asteraceae coincide with marked shifts in transposable element abundance. BMC Genomics 16:623.

Steinbiss S, Willhoeft U, Gremme G, Kurtz S (2009) Finegrained annotation and classification of *de novo* predicted LTR retrotransposons. Nucleic Acids Research 37:7002– 7013

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673–4680

Vitte C, Panaud O (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenetic and Genome Research 110:91–107

Weber B, Heitkam T, Holtgräwe D, et al. (2013) Highly diverse chromoviruses of *Beta vulgaris* are classified by chromodomains and chromosomal integration. Mobile DNA 4:8

Wei F, Stein JC, Liang C, et al. (2009) Detailed analysis of a contiguous 22-mb region of the maize genome. PLoS Genetics 5:e1000728

Wu F, Mueller LA, Crouzillat D, Pétiard V, Tanksley SD (2006) Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics 174:1407–1420

Table 1 Number and location of nucleotide variants in the *Marshallia obovata Del/Tekay* element family. NonORF indicates the noncoding region between the 3′ end of the LTR1 and the start of the gag-pol region. ORF indicates SNPs that occurred inside the gag-pol region but outside of conserved domains.

Region	No.	Sub-Region	No.	Syn.	N.Syn
Non-coding	29				
		LTR	26		
		NonORF	3		
Coding	21				
		ORF	13	5	8
		Chromo	1	0	1
		RNase_H	2	2	0
		RVP_2	2	1	1
		RVT_1	1	1	0
		UMSBP	2	1	1
Total	50	All	50	10	11

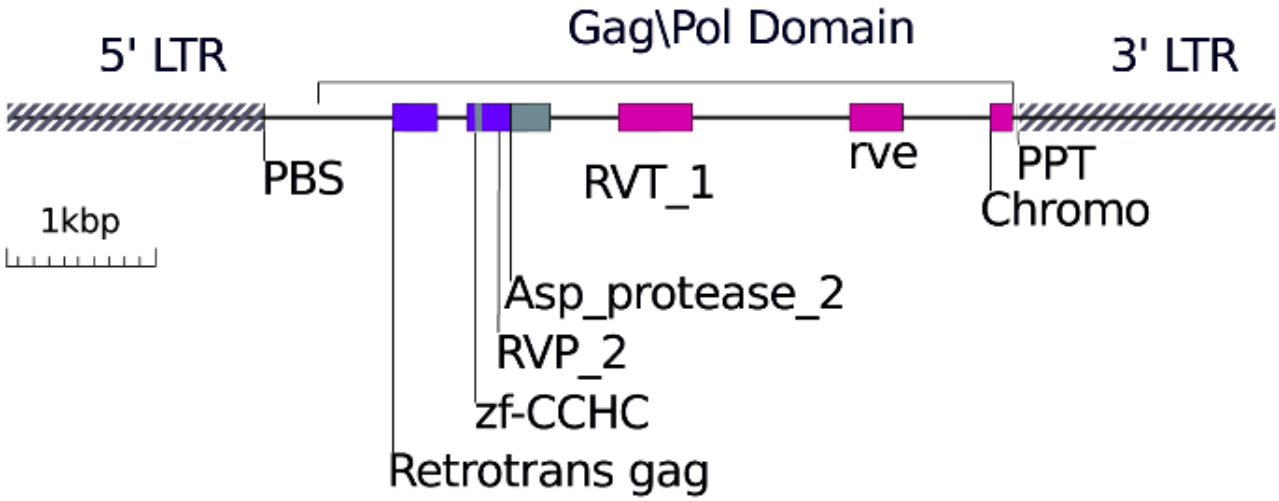


Figure 1 Schematic of *Marshallia obovata Del/Tekay* element with sequence-based annotation of characteristic retroelement features: 5′ LTR (1–1675), primer binding site (PBS, 1681–1689), Retrotransposon gag protein (Retrotrans_gag, 2522–2821), zinc finger (zf-CCHC, 3065–3113), retroviral aspartyl protease (RVP_2, 3041–3625), aspartate protease (or pepsin-like aspartate proteases, Asp_protease_2, 3293–3559), reverse transcriptase (RVT_1, 4007–4486), Ribonuclease H (RNase_HI_RT_Ty3, 4766–5110), integrase core domain (rve, 5522–5878), chromodomain (Chromo, 6437–6586), polypurine tract (PPT, 6615–6631), and 3′ LTR (6634–8308)


Galadriel
M. obovata
Del/Tekay
Chromoviruses
Fungi and
Vertebrate
Reina
CRM
Micropia/Mdg3
CsRN1
Gmr1
Tat
Mag

Figure 2 Maximum likelihood tree from analysis of RVT domain in *Ty3/Gypsy* elements with emphasis on chromoviruses

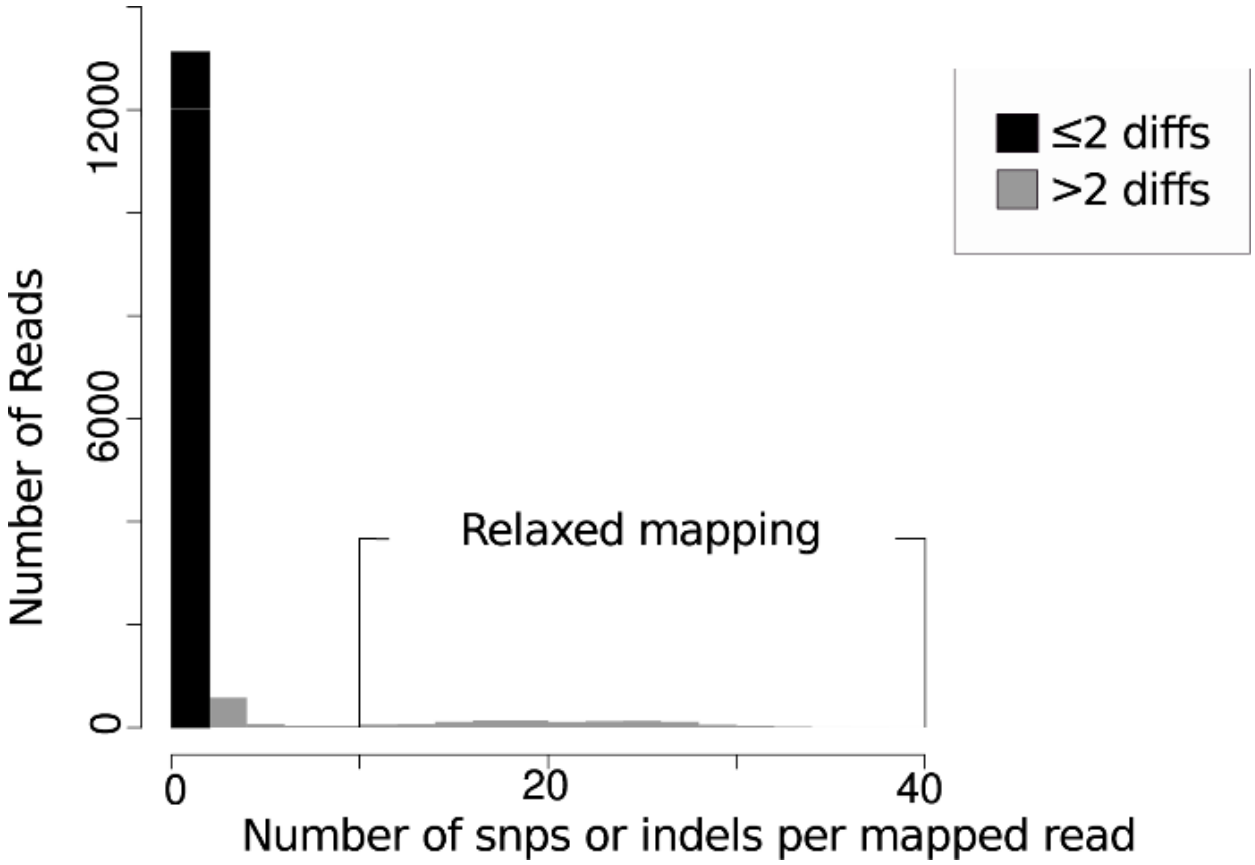


Figure 3 Number of SNP/indel variants (diffs) in each read that mapped to the reference

Marshallia obovata Del/Tekay element with relaxed mapping parameters.

Chapter 4 An investigation of interspecies relationships within *Marshallia*

Abstract

Marshallia schreb. [Asteraceae] is a genus of southeastern endemic wildflowers, that contains eight species of uncertain phylogenetic relationship, one of which, *M. mohrii* Beadle & F.E. Boynt., is Federally Listed and an alltetraploid of unknown origin. Sequencing, assembly, and phylogenetic analysis was undertaken for 17 samples from within *Marshallia* to resolve basic species to species relationships. All samples were sequenced using the Illumina platform, and assemblies of the plastid, ribosomal DNA, and mitochondrial sequences were produced. Phylogenetic analysis of ribosomal DNA seqeuences provided strong support for currently described species and placed *M. mohrii* sister to *M. trinervia* (Walter) Trel.. Phylogenetic analysis of plastid and mitochondrial sequences revealed inter-species cytoplasmic sharing. Read clustering analysis of the 17 samples revealed that there were differences in transposable element superfamily abundance among species within *Marshallia*. Finally, eight samples of *Marshallia* transcriptomic DNA were sequenced, assembled, and used to further investigate the parentage of *M. mohrii*. Orthologs from the transcriptomic assemblies were used in supermatrix and gene tree phylogenetic analyses. Both approaches suggested that *M. caespitosa* Nutt. ex DC. and *M. ramosa* Beadle & Boynt. were possible parents of *M. mohrii* in

addition to *M. trinervia* which was identified by the phylogenetic analysis of ribosomal DNA. Another possible interpretation of these results is that the other parent of *M. mohrii* is extinct. More work will need to be done to determine the parentage of *M. mohrii*.

Introduction

The genus *Marshallia Schreb.* [Asteraceae] occurs throughout the Appalachians, Cumberland Plateau, and Southeastern Coastal Plain east of the Mississippi and from Texas to Kansas west of the Mississippi River (Channell 1957). The genus contains eight species (Watson and Estes 1990; Weakley and Poindexter 2012), with seven species recognized by Watson and Estes (1990): *M. caespitosa* Nutt. ex DC., *M. obovata* (Walter) Beadle & F.E. Boynt., *M. grandiflora* Beadle & F.E. Boynt., *M. graminifolia* (Walter) Small, *M. mohrii* Beadle & F.E. Boynt., *M. ramosa* Beadle & F.E. Boynt., *M. trinervia* (Walter) Trel., and *M. legrandii* Weakley. The eighth and most recent species was described from two very restricted populations (Weakley and Poindexter 2012). Watson and Estes (1990) divided *Marshallia* into four complexes in keeping with the number laid out by Channell (1957) in his revision of the genus: complex 1 (*M. grandiflora, M. trinervia,* and *M. mohrii*), complex 2 (*M. caespitosa*, *M. obovata*), complex 3 (*M. graminifolia*) and complex 4 *(M. ramosa*). Recent work by Hansen and Goertzen (2014), using the internal transcribed spacer (ITS), clusters *M. legrandii* with complex 1 by placing *M. legrandii* sister to *M. grandiflora*. All species exhibit a range of morphological characters that overlap to some degree (Beadle and Boyton 1901; Channell 1957; Watson and Estes 1990; Weakley and Poindexter 2012) and phenetic analysis of a set of 38 morphological characters suggests that a combination of traits is sufficient for determining differences among the lineages examined. Each of the species appears to be separated from other congenerics by divergent

habitat (Channell 1957) or flowering phenology, as in the case of *M. legrandii* and *M. obovata* (Weakley and Poindexter 2012). Of the seven species recognized by Watson and Estes (1990), all exhibit a degree of allogamy with the exception of *M. ramosa* which shows signs of chromosomal inversions and exhibits multivalent chromosome structures in pollen mother cells (Watson and Estes 1987, 1990). Of particular interest is the Federally-listed allotetraploid *Marshallia mohrii* (U.S. Fish and Wildlife Service1988). Its distribution is extremely restricted with 22 known locations reported in its Recovery Plan. It typically occurs in wet acidic soil in habitats maintained by fire or regular mowing (Norquist 1991). *Marshallia mohrii* has been traditionally grouped within complex 1 (with *M. grandiflora* and *M. trinervia*) because it is similar in leaf structure and flowering habit to *M. grandiflora,* and because *M. mohrii* occurs with the ranges of *M. grandiflora* and *M. trinervia* (Channell 1957; Watson and Estes 1990).

The aim of this study is to examine interspecies relationships within the genus *Marshallia* by leveraging two cost-efficient sequencing methods, genome skimming and *de novo* transcriptome assembly. Genome skimming is the sequencing of low coverage whole genomic DNA, and it is particularly useful for the assembly of organellar genomes and the assembly of highly repetitive sequences from the nuclear genome (Bock et al. 2014). These datasets provide a natural sample of the genomic sequence diversity for use in the estimation of repetitive element abundance (Staton and Burke 2015), the assembly of the 26S, 5.8S, and 18S ribosomal DNA cistron (rDNA) called ultra-barcoding (Bock et al. 2014; Oscar et al. 2017), and the assembly of chloroplasts and partial mitochondrial sequences (Straub et al. 2012; Grass Phylogeny Working Group II 2012; Malé et al. 2014; Vargas et al. 2017). The assembly of transcriptomes is also recognized as a way to sequence genes from species with no published genomes and has been

used successfully to resolve complicated phylogenetic relationships (Francis et al. 2013; Wickett et al. 2014; Wang et al. 2017).

Materials and Methods

Sequencing and Read Filtering

DNA extraction and sequencing is detailed in Hansen and Goertzen (2014): briefly, whole genomic DNA was extracted using modified CTAB (Doyle and Doyle 1987) or an E.Z.N.A ® Plant DNA Kit (Omega Bio-tek, Inc., Norcross, GA) from leaves of specimens that were collected and vouchered in Auburn Herbarium (AUA), or grown from seeds maintained in greenhouses. Whole genomic DNA was submitted for sequencing to HudsonAlpha Institute for Biotechnology (Huntsville, AL) and was sequenced as paired-end reads with 100 bp of sequencing at the 5′ and 3′ end of each fragment. RNA was extracted from fresh tissue using Plant RNA extraction kit (Qiagen, Hilden, Germany) per manufacturer's protocol or fresh tissue was submitted to Auburn University Genomics and Sequencing Laboratory (Auburn, AL) for in-house RNA extraction. Sequencing libraries were prepared using Illumina mRNA TruSeq kit and sequenced on ILLUMINA HISEQ 1500 platform at Auburn University Genomics and Sequencing Laboratory.

SRA accessions for *Inula racemosa* Hook.f. (SRR5237205), *Inula ensifolia* L. (SRR5237206), *Marshallia caespitosa* (SRR5237261), *Helenium autumnale* L. (SRR1141042), *Gaillardia pulchella* Foug. (SRR5237268), *Gaillardia aristata* Prursh (SRR5237184), *Glebionis segetum* L. Fourr. (SRR797217), *Guizotia abyssinica* (L.f.) Cass. (SRR5237269), and *Guizotia*

scabra (Vis.) Chiov. (SRR797218) were downloaded from NCBI SRA (https://www.ncbi.nlm.nih.gov/sra/ 28 September 2017).

All paired end reads were processed using AfterQC v0.9.3 (Chen et al. 2017), single end reads were processed using cutadapt v1.7.1 (Martin 2011) for 454 sequences or for single end illumina sequences, trim_galore v0.4.0 (Krueger 2015), a pipeline which implements cutadapt v1.7.1 (Martin 2011), was used for in house pipelines and trimmomatic (Bolger et al. 2014) was used for transcriptome assembly (Table 2).

Genome Skimming rDNA Assembly and Analysis

Full length copies of the ribosomal cistron were assembled through targeted assembly, and validation for all genome skimming datasets and consensus sequence were called for cDNA read sets. To create genome skimming assemblies, ribosomal contigs were extracted from a preliminary shotgun assembly of unfiltered reads for each dataset using the ribosomal subunit sequences downloaded from NCBI (*Artemsia absinthium* L. (EU649668.1), *Helianthus annuus* L. (DQ865267.1), *Ixeridium cf.* (AB766222.1), and *Krigia wrightii* (A. Gray) K.L. Chambers ex Kim (L20480.1)) to identify candidate contigs. A map of candidate contigs was produced with Bowtie 2 v2.2.9 (Langmead and Salzberg 2012) for each dataset; mapped reads and their pairs were extracted using SAMtools v1.2 (Li et al. 2009), assembled with Ray v2.3.2 (Boisvert et al. 2010) then iteratively extended with Price TI v1.0.1 (Ruby et al. 2013), joined with cap3 version date: 08/06/13 (Huang and Madan 1999) or, in recalcitrant cases, joined manually. Contigs joined manually had a minimum overlap of 20 bp. Assembled sequence was determined complete when 5′ and 3′ ends both contained highly conserved non-transcribed promoter region

TATAGGGGG (Linder et al. 2000) or circularized, and assembly was validated by even depth of concordant mapped reads in Bowtie 2. Final sequence for analysis was determined using pysam_consensus.py (https://goo.gl/LGM6iu) using pysam v0.13.0 (http://pysam.readthedocs.io/en/latest/api.html: Accessed Aug. 1, 2017). Sequences were annotated by comparison to a confamilial using blast+ NCBI (NCBI Accession: KF767534.1) (Camacho et al. 2009; Bock et al. 2014). Internal transcribed spacers were validated by comparison to published secondary structural models (Goertzen et al. 2003; Xu et al. 2011). Assembled sequences were placed into alignment using mafft v7.221 (Katoh and Standley 2013) with flag --max-iterate 1000, visualized with SeaView (Gouy et al. 2010), and gappy columns were removed using trimal v1.2 (Capella-Gutiérrez et al. 2009) using the flag -automated1. A maximum likelihood tree was created with RAxML v8.2.9(Stamatakis 2014) and visualized with FigTree v1.4.1 (Rambaut 2009).

Genome Skimming Plastid Assembly and Analysis

Plastid assembly was accomplished in two steps: first, the assembly of and annotation of a gold standard references and, second, the use of the gold standard to target assembly to the chloroplast genome. Reads were cleaned using Sickle Pe (Joshi NA 2011) and assembled into preliminary contigs using Ray. Plastid fragments were extracted from these contigs using *Helianthus annuus* and *Guizotia abyssinica* (NCBI Accessions: NC_007977.1, NC_010601) as references. Fragments were iteratively extended by mapping reads to plastid contigs, extracting reads and mapped and unmapped pairs with SAMtools and assembling new expanded sets of reads using Ray. When fragments overlapped by more than 20 bp they were joined together.

Poly-N-regions were verified using majority rule of reads that completely transversed Poly-N-region and were anchored on both the 5′ and 3′ sides by conserved sequences. Boundaries between the long single copy (LSC) region, inverted repeats (IR), and short single copy region (SSC) were validated using targeted assemblies of reads that mapped concordantly across each boundary: IRb/LSC LSC/IRa, IRa/SSC SSC/IRb. Annotation was carried out using BLAST+, dogma (Wyman et al. 2004), and Artemis v16.0.0 (Rutherford et al. 2000).

To assemble all genomes, a reference map was made from the initial gold standard assembly: it contained the large single copy region (LSC), short single copy region (SSC), and 1 copy of the inverted repeats (IR). Cleaned reads (Table 1) from each set were mapped to reference with Bowtie 2. Reads were then extracted by region and assembled using Ray and Price TI. Regions were connected manually so that they could be assembled in the same orientation and so that problem regions could be fixed with further assembly using Ray, Price TI, and manual extension if required. Maps of each genome were produced using Bowtie 2 and the GATK v3.6 (McKenna et al. 2010) best practice pipeline to realign indels (Van der Auwera et al. 2013). Each map was inspected in tablet and used to call a consensus. Consensus sequences were inspected using tablet and modified in low quality mapping regions (https://goo.gl/LGM6iu). Poorly assembled regions were scaffolded together using paired end reads ambiguity code N and, if no connecting read could be found, contigs were scaffolded based on alignment with the original reference. Final copies of contigs were circularized through by the addition of IR to the 3’ end of each assembly and validated by a final mapping of concordant reads using Bowtie 2 and flags --no-mixed, --no-discordant and --no-unal. The originally assembled *Marshallia obovata* (M3) was annotated using Dogma and Artemis, and these annotations were transferred to subsequent *Marshallia* assemblies using Plann (https://github.com/daisieh/plann; Accessed:

Nov 1, 2017) (Huang and Cronk 2015). During annotation poly-N regions occuring within proteins were modified to be inframe if they were not originally.

The LSC, IR, and SSC were extracted from each annotated plastid genome using extractseq from Emboss v.6.6.0.0 (Rice et al. 2000), aligned with mafft with flag --max-iterate 1000, visualized, and manually adjusted with SeaView. Trimal was used to eliminate columns in which 20% or more of the characters were gaps. RAxML was run on alignment using GTR+Gamma model with 1000 bootstraps.

Genome Skimming Mitochondrial Assembly and Analysis

A pseudo-molecule approach was employed to create mitochondrial sequences for alignment, because plant mitochondrial genomes have exceptionally high rates of recombination and rearrangement that make the assembly of a full length mitochondrial genome untenable with paired end libraries (Vargas et al. 2017). Putative mitochondrial contigs from each shotgun assembly were identified using CDS sequences from *Diplostephium hartwegii* (NCBI Accession: NC_034354.1) as queries in a blast search. A set of representative contigs was chosen from the sample with the most complete mitochondrial fraction. Contigs were modified to contain no repeats greater than 300 bp, regions small enough to be traversed by concordant read pairs from the paired end libraries. Here two different assemblies were made: the first assembly we followed the protocol used in our chloroplast assembly, and the second assembly we created from a consensus based approach. For the *de novo* assembly, cleaned reads were used to target each unique contig and assemble it, and mapping was then used validate assembly and remove

any contigs of chloroplast in conjunction with BLAST+. Assembled sequences were combined into pseudo-molecules using blast and custom python script.

For the consensus approach, the unicontig set was tested for chloroplast contaminants using chloroplast reference and chloroplast contigs were excluded. A concordant competitive mapping was performed using Bowtie 2 in which only reads that mapped correctly in pairs to the mitochondrial genome and not the chloroplast genome were considered. The reference contained all assembled ribosomal DNA and plastid DNA along with the mitochondrial contigs, this ensured the exclusion of chloroplast or ribosomal reads which might otherwise improperly map to the mitochondria and obfuscate the mitochondrial sequence by virtue of their abundance.

This approach evened out the number of mapped reads across all chloroplast transfer regions shorter than the fragment length of the paired end library, and prevented the plastid read contamination of mitochondrial sequence. Maps were made with Bowtie 2 and were filtered with cigar_filter_v2.py --paired (https://goo.gl/hq9dK6) to exclude speciously mapped reads and their mates. Indels were realigned using GATK best practice pipeline (McKenna et al. 2010; Van der Auwera et al. 2013) and consensus was called using pysam_consensus, and concatenated using FASconCAT.

Preliminary alignments were produced with mafft using default parameters and cleaned manually with SeaView to exclude gaps and compared using Paup* v4.0b10 (Swofford 2003). Both matrices produced maximum parsimony trees with the same topologies, but the consensus sequence approach provided more high quality sequence (395 parsimony informative characters) than the contig assembly approach (146 parsimony informative characters). New alignments were created from consensus sequences, using mafft -max-iterate 1000; columns containing gaps

were excluded with trimal and all sequences were concatenated using FASconCAT. RAxML was run on the resulting supermatrix with model GTR+Gamma with 1000 bootstraps.

Transcriptomic, rDNA, and Organellar Sequences

To create a set of coding sequences and ribosomal DNA for comparison, cDNA reads were run through a mapping filtering pipeline. Briefly, reads were competitively mapped against the target reference sequence, rRNA cistron *M. trineriva* M2, and unjoined CDS sequence was extracted from reference *Marshallia* plastid M3 genome and unjoined CDS mitochondrial sequence from *Diplostephium hartwegii* (NCBI Accession: NC_034354.1). Mapped reads were cleaned with trim_galore --illumina flag, and then remapped to reference sequence using Bowtie 2 using local flags and soft clipped reads were filtered using cigar_filter.py (https://goo.gl/hq9dK6). Indels were realigned using GATK best practices (Van der Auwera et al. 2013) with picard tools and GATK (McKenna et al. 2010), and consensus was called using pysam_consensus.py (https://goo.gl/LGM6iu).

Ribosomal DNA was combined with previously assembled sequences aligned with mafft --max-iterate 1000, and visualized and trimmed with SeaView, after which gappy columns were removed with trimal -automated1 flag. The alignment was analyzed using RAxML with GTR+ Gamma parameters with 1000 bootstraps. Coding sequences for the mitochondrial and plastid genomes were combined with assembled sequences, aligned with mafft --max-iterate 1000, and then trimmed with trimal to remove columns composed of 25% gaps. Sequences were concatenated using FASconCAT and then Partition-Finder with --raxml flag was run. RAxML was run on the cleaned alignment using a GTR + Gamma model and a 1000 bootstrap analysis.

After the first RAxML run, in which coding organellar sequences were run together, it became apparent that sequences were segregating based on their nucleic acid of origin and not the organism, even after further manual alignment revision. Sequences assembled from genomic DNA were removed and RAxML was rerun with the same parameters.

Transcriptome Assembly, Annotation, and Phylogenetic Analysis

Paired end datasets for *Marshallia mohrii* (M20, M21), *M. graminifolia* (M39), *M. ramosa* (M38), *M. caespitosa* (M10), *M. grandiflora* (M29), *M. obovata* (M3), *M. trinervia* (M2)*, Helenium autumnale* (NCBI Accession: SRR1141042), *Glebionis segetum* (NCBI Accession: SRR797217), and *Guizotia abyssinica* (NCBI Accession:SRR5237269) were cleaned using afterQC, and single end reads were cleaned using trimmomatic (Bolger and Giorgi 2014) built into Trinity v2.4.0 (Grabherr et al. 2011); all read sets were assembled with Trinity, proteins and CDS sequences were predicted with TransDecoder (Haas and Papanicolaou 2016), and annotations were performed using Trinotate pipeline (Hébert). A set of unicontigs was obtained for all protein sequences with CD-HIT v4.7 (Fu et al. 2012) run with default parameters. A set of single to low copy genes was determined using OrthoFinder v1.1.8 (Emms and Kelly 2015) and in-house pipeline orthogroup_extract.py. Briefly, orthogroup extract allows for the detection of putative single copy genes that fall below OrthoFinder's default cutoff (Katoh and Standley 2013). Trimal -automated1 flag was used to trim genes and custom python script was used to remove sequences before a 10 nt gap in the first 50 nt of alignment or after a 10 nt gap in the last 50 nt of alignment. Sequence order was standardized and placeholder taxa were added using fasta_ghost.py (https://goo.gl/dcV1gz). Gene alignments were concatenated

using FASconCat.pl. Partitions were created using Partition-Finder with --raxml flag and trees were created in RAxML with GTR + Gamma model.

A subset of single low copy genes was extracted using orthogroup_extract.py. These sequences were aligned and iteratively cleaned. The alignment cleaning pipeline is described in the following sentences. General alignment with mafft default parameters, infoalign from Emboss v.6.6.0.0 (Rice et al. 2000), was used to determine poor alignments and sequences with greater than 10 percent divergence from the alignment consensus not including gaps. The 10 percent cutoff was determined by viewing several alignments by eye with SeaView. Passing sequences were retrieved with select_contigs.pl (White 2009). Trimal was used to reduce gaps with -automated1 flag. Custom python script was used to remove all sequence prior to last 10 bp gap within the 5' half of the alignment or after the earliest 10 bp gap in the 3' half of the alignment to remove any poorly assembled or chimeric ends of the assembled transcript. Finally, columns containing gaps were removed using trimal to create 2 different matrices: relaxed, which allowed up to 2 gaps per column, and strict, which allowed no gaps. Cleaned alignments were concatenated with FASconCAT, partitions were created using Partition-Finder with --raxml flag, and RAxML was used to generate best maximum likelihood tree for the supermatrix. Astral-III (Zhang et al. 2017) was run on all cleaned genes individually to determine putative parents for *Marshallia mohrii.* Dendropy v4.3.0 package for python 3 was used to determine the nearest non-target taxa to *M. mohrii* using unweighted patristic distance.

Assessment of the Repetitive Genomic Fraction

Here a two pronged approach was used: 1) changes in abundance for a single previously described *Del/Tekay* element were tracked, and 2) changes at the level of superfamily were tracked. To assess if there was a change in copy number of previously described Gypsy element in *Marshallia* (Hall and Goertzen 2016), the element was downloaded and all raw reads were mapped against the sequence with Bowtie 2 using the --qc-filter flag along with end to end alignment of reads. Final number of reads was compared to overall number to derive the proportion of sequenced data represented by reads mapping to the *Del/Tekay* element. To assess changes in superfamily abundance, reads were cleaned with AfterQC and filtered using Bowtie 2 and SAMtools to exclude any read pair for which one of the reads mapped to the previously assembled ribosomal DNA, mitochondrial contigs, or plastid genomes. Whole datasets were too large for Transposome v0.10.0 (Staton and Burke 2015), so each sample was randomly downsampled to 400 K reads three times using a modified version of an online code snippet, with a modification to the random sampling parameter so that it sampled without replacement to avoid artificially inflating the values for any given sequence (https://goo.gl/fqMVYZ accessed June 22, 2017). Even with initial downsampling, some samples produced data structures too large to be analyzed with 256 Gb of memory. These samples were further downsampled to 50 K using the bash utility head after trying several larger values. The number of reads used to identify a repetitive cluster was decreased to 25 from 100 for reduced datasets. Repetitive clusters were annotated using a custom repeat database made by running RepeatMasker v3.2.7 (Smit et al. 2017) on a preliminary shotgun assembly of *M. obovata* (M3). Repeats were extracted using SAMtools. Results of Transposome analysis were analyzed for clustering with principal component analysis using package scikitlearn (Pedregosa et al. 2011), pandas (McKinney 2010), and matplotlib (Hunter 2007) for python 3.4.3 (Van Rossum 2014). Additionally, two ANOVAs

which controlled for the repeated measures were run with R (v3.3.3), using packages aov, nlme (Pinheiro et al. 2017), and multcomp (Hothorn et al. 2009), to determine if there was any variation in total proportional repeat abundance and to determine if there was any variation in the difference between the two most abundant superfamilies (Copia and Gypsy). After finding significance for relative difference between Copia and Gypsy, a post hoc test was performed using Tukey's HSD with package multcomp.

Results

Alignment and Comparison of Genomic and Transcriptomic rDNA and Organellar Sequence

Assembled plastid and rDNA sequences have been accessioned at NCBI (MH037167, MH037168, MH037169, MH037170, MH037171, MH037172, MH037173, MH037174, MH037175, MH037176, MH037177, MH037178, MH037179, MH037180, MH037181, MH037182, MH037183, MH037184, MH037185, MH037186, MH037187, MH037188, MH037189, MH037190, MH037191, MH037192, MH037193, MH037194, MH037195, MH037196, MH037197, MH037198, MH037199, MH037200). Both rDNA trees produced congruent trees and clustered samples to create monophyletic species with the exception of *Marshallia trinervia* and *M. mohrii* in the combined transcriptomic genomic rDNA tree. These sequences are known to cluster closely together (Hansen and Goertzen 2014). The rooted ribosomal tree shows a bifurcation at the base of *Marshallia* creating a *M. graminifolia-M. grandiflora* clade and a *M. obovata-M. caespitosa* clade (Fig. 1). Genomic rDNA had 604 distinct alignment patterns and 4.18% undetermined characters, with a tree length of 0.084364

and alpha parameter of 0.02. The genomic rDNA tree possessed high bootstrap support (Fig. 2). Alignment of the whole plastid produced had 377 distinct alignment patterns, with 0.05% undetermined characters (Fig. 3). The best tree had a total length of 0.007122 and an alpha parameter of 0.02. Mitochondrial pseudo-molecule alignment produced 352 distinct alignment patterns and the best tree had a total tree length of 0.006727 with an alpha parameter of 0.2, and a tree topology highly congruent with the whole plastid tree, but it had longer internal internodes than the plastid tree (Fig. 4). It is apparent from these two trees that there are distinct cytotypes that are not constrained by species identity and are incongruent with rDNA. The rooted tree made with the coding sequences of both the plastid and the mitochondrial genome show that the cytotype found in *M. trinervia* (M2) is basal with respect to the other transcriptomic samples (Fig. 5).

Transcriptome Assembly and Analysis

Trinity assembly produced a range of unique genes with minimum of 41 K and a maximum 173 K (Table 3). OrthoFinder found a total of 28,023 orthogroups accounting for 92.3% of the 523,641 sequences that were consolidated by CD-HIT. Eighty-five orthogroups contained a single copy of all species; orthogroup_extract.py was used to query OrthoFinder results to exploit the mostly complete orthogroups identified by orthorfinder and return nucleotide sequences for identified genes. One hundred and seventy-three single copy genes with outgroups were concatenated into a 123,122 nt supermatrix. Partition-Finder returned 15 partitions which were used with RAxML which returned alpha parameters ranging from 0.20 to

0.691948 and a total of 26,527 alignment patterns and 10.16% missing data or gaps. The best tree has a length 0.678941 (Fig. 6).

For single copy genes within *Marshallia* the sequence variance was too low for a successful Partition-Finder run, tested multiple times on the strict matrix. Sequences were treated as 1 partition in light of repeated failure due to low sequence divergence. Tree topology is conserved between strict and relaxed matrix and tree lengths are comparable with higher bootstrap support for the nodes in the tree produced by RAxML from the relaxed matrix (Fig 7). The relaxed matrix contained 885 aligned sequences 536,368 nt in length with 4,346 distinct alignment patterns, an alpha parameter 0.02, 9.74% gaps or missing, and a tree length of 0.045346. The strict matrix contained 424 aligned sequences 23,9476 nt in length with 939 distinct alignment patterns, alpha parameter 0.02, and best tree length of 0.042048. Astral placed *M. mohrii* sister to *M. ramosa* and assigned low support at all nodes (Fig. 8). Tallies of taxa with minimum patristic distance relative to *M. mohrii* showed that *M. graminifolia* (M39 122), *M. grandiflora* (M29 138), and *M. obovata* (M3 158) had the lowest tallies while *M. trinervia* (M2 238), *M. ramosa* (M38r 293), and *M. caespitosa* (M10r 298) had the highest tallies.

Analysis of the Repetitive Fraction Using Low Coverage Reads

The *Del/Tekay* element *Marobo* was present in all datasets and occurred across a range of percentages, from 0.0172 (*M. trinervia)* to 0.0936 (*M. grandiflora*). There was no significant difference in the total proportion of the repetitive fraction of the genomes which have a grand mean of 0.6878 (min=0.5703, max=0.7365): see Table 6 for abundance by subsample. Since there was no statistical difference among all groups regardless of sampling depth, a comparison

of the relative abundance of Copia compared to Gypsy elements within each group was undertaken. Here significant differences were found between the relative amount of Copia and Gypsy elements. *Marshallia obovata* and *M. grandiflora* both exhibited significantly more balanced Copia and Gypsy superfamily fractions than their congenerics (Fig. 9). Principal component analysis accounted for more 90% of the variation within the data for the first two principal components. Graphing of the transformed data showed little cohesion among groups; for example *M. mohrii* occurred at either end of the main cluster. While it is clear that there is some resolution among individuals, there are no pure species clusters, with the possible exception of *M. legrandii*, however, since only one sample of *M. legrandii* was available for this analysis it is premature to suggest that *M. legrandii* clusters as a species (Fig. 10).

Discussion

These analyses show a pattern consistent with the species of *Marshallia* diverging relatively slowly. The repetitive fraction of the genomes analyzed displays some nuanced variation. The plastid and mitochondrial genes and genomes are relatively homogeneous to the point that they appear to be interchangeable, yet phylogenetic trees constructed from rDNA suggest that the nuclear genomes are consistent within previously morphologically circumscribed species, and genes assembled from the transcriptomes demonstrate a slow rate of divergence.

Considering the variation of the *Del/Tekay* element within *Marshallia,* in tandem with the Transposome results, suggests that there are detectable differences among individuals within species with regard to transposable element content. This is particularly fascinating for *Marshallia mohrii* (an allopolyploid) and *M. caespitosa,* which includes autopolyploid and

diploid populations. The effects of genomic shock are often studied in neoplolyploid systems such as crop plants and experimental hybrids, however, little work has been done on the persistence of genomic shock effects within meso-polyploid systems (Soltis et al. 2016) such as *M. caespitosa* and *M. mohrii*. These species offer a chance to test mesoscale evolutionary predictions of genomic shock on a wildflower system. It also appears that some force, such as recombination (Devos et al. 2002; Michael 2014), is acting differentially among species to reduce the ratio of Gypsy to Copia elements within *Marshallia*.

The samples collected for *Marshallia* exhibit distinct cytotypes. Given the short internodes within the plastid tree it is not surprising that there is some uncertainty in the placement of *M. obovata* sample M32. Furthermore, the pattern of cytotypes does not match the pattern of either the rDNA or the transcriptome analysis. This state strongly suggests a history of introgression among *Marshallia* species. Cytoplasm capture can occur under a variety of conditions, including when the plastid and mitochondrial sequence represents neutral loci or when a cytotype confers a specific advantage (e.g. cytonuclear interactions) (Tsitrone et al. 2003). Given the relatively few differences among plastid and mitochondrial sequences, it seems more likely that interspecific cytotypic spread is the product of introgression and drift as opposed to a selective sweep. In either case, the resulting cytotype can sweep the population in the presence of limited to modest allogamy. Simulations show that a modest amount of gene flow between populations during an invasion event can lead to near fixations within the invading population with as little as 3% outcrossing success, and only slightly higher values (*ca*. 10%) can lead to similar results in the case of prolong sympatry (Currat et al. 2008). *Marshallia* are known to be allogamous with modestly high outcrossing rate, which has a grand mean of 15% (Watson and Estes 1990). This is more than enough outcrossing to allow for the transfer of plastids,

mitochondria, and neutral alleles among species, and it strongly suggests that populations of these species were sympatric at early stages of divergence even if they are currently isolated and relegated to separate habitats (Channell 1945; Watson and Estes 1991).

The rDNA analysis provides the clearest signal that the eight described species within *Marshallia* represent monophyletic lineages. The branching order of these lineages does not follow the pattern expected when compared with Channell's revision or more recently the phenetic clustering of morphological characters applied by Watson and Estes (1990). Admittedly, the ribosomal cistron used for this analysis is essentially one locus that undergoes concerted evolution within lineages (Hillis and Dixon 1991), yet its complete support of the described species, with the exception of *M. trinervia* and *M. mohrii,* adds additional weight to the observation of cross species cytotypic sharing. Because the rDNA from the same samples does not follow the cytotypic topology, misidentification can be ruled out as the reason for the strange cytotypic topology.

It is all the more noteworthy that the phylogenetic analysis of the transcriptomic sequences does not conform to expected species complexes (Channell 1957; Watson and Estes 1990), or the phylogenetic signature of the ribosomal DNA. Salient points of contrast are the change in position of *M. obovata* and *M. trinervia*. Within the ribosomal tree, *M. graminifolia* and *M. grandiflora* form the basal node sister to *M. obovata*, *M. trinervia*, *M. mohrii, M. ramosa,* and *M. caespitosa,* in marked contrast to the transcriptomic tree in which *M. trinervia* is sister to all *Marshallia,* and *M. obovata* is sister to *M. grandiflora* and *M. graminifolia*. Finally, this tree topology places *M. mohrii* sister to *M. caespitosa* and *M. ramosa*. The close molecular relationship of *M. mohrii* to both *M. trinervia* and the *M. caespitosa-M. ramosa* complex is interesting for at least two reasons. First the close relationship of *M. trinervia* and *M. mohrii* is

expected given their proximity and Channell's (1957) and subsequently Watson and Estes' (1990) Grandiflora species complex, but this along with work by Hansen and Goertzen (2014) is some of the first solid evidence that *M. trinervia* is a likely genome donor. Second the close relationship of *M. mohrii* to the *M. caespitosa*-*M. ramosa* clade is unexpected but may provide the key to identifying the second genome donor. It seems unlikely that the second genome is from *M. ramosa* since *M. mohrii* does not also possess the characteristic chromosome inversions (Watson and Estes 1990). The relationship of *M. mohrii* to *M. caespitosa* echoes the clustering of these two species found in the jackknife analysis of canonical variance related to morphological characters. These lines of evidence rule in the possibility of either *M. caespitosa* or, in a less likely scenario, *M. ramosa* as parental genome donors for *M. mohrii*. However they do not necessarily rule out *M. grandiflora* which also clustered with *M. mohrii* during the same jackknife analysis, and also possesses a limited amount of support among the gene trees. *Marshallia mohrii* is connected to *M. ramosa, M. caespitosa,* and *M. trinervia* through transcriptome analysis. It is also clearly related to *M. obovata* by cytotype. Taken together these data suggest a set of slowly diverging lineages at the level of species with a history of introgression.

References

Beadle CD, Boyton FE (1901) Revision of the species of *Marshallia*. Biltmore Bot 1:3–10

Bock DG, Kane NC, Ebert DP, Rieseberg LH (2014) Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. New Phytol 201:1021–1030

Boisvert S, Laviolette F, Corbeil J (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol 17:1519–1533

Bolger A, Giorgi F (2014) Trimmomatic: a flexible read trimming tool for illumina NGS data.

URL http://www usadellab org/cms/index php

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Camacho C, Coulouris G, Avagyan V, et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973

Channell RB (1957) A revisional study of the genus *Marshallia* (Compositae). Contributions from the Gray Herbarium of Harvard University 41–130

Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res 12:1075–1079

Doyle JJ, Doyle JL (1987) CTAB DNA extraction in plants. Phytochemical Bulletin 19:11–15

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157

Francis WR, Christianson LM, Kiko R, et al (2013) A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. BMC Genomics 14:167

Fu L, Niu B, Zhu Z, et al. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152

Goertzen LR, Cannone JJ, Gutell RR, Jansen RK (2003) ITS secondary structure derived from comparative analysis: implications for sequence alignment and phylogeny of the Asteraceae. Mol Phylogenet Evol 29:216–234

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221–224

Grabherr MG, Haas BJ, Yassour M, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

Grass Phylogeny Working Group II (2012) New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. New Phytol 193:304–312

Haas BJ, Papanicolaou A (2016) TransDecoder (find coding regions within transcripts)

Hall ND, Goertzen LR (2016) Sequencing and characterization of the *Del/Tekay* chromovirus family in *Marshallia obovata* (Asteraceae). Paysonia 5:1–7

Hansen CJ, Goertzen LR (2014) Validation of nrDNA ITS as a DNA barcode for *Marshallia* (Asteraceae). Paysonia 3:5–10

Hébert FO Trinotate_pipeline: Annotation Pipeline-Trinotate. Zenodo. 2016

Hillis DM, Dixon MT (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. Q Rev Biol 66:411–453

Hothorn T, Bretz F, Hothorn MT (2009) The multcomp package. Technical Report 1.0-6, The R Project for Statistical Computing, www. r-project. org

Huang DI, Cronk QCB (2015) Plann: A command-line application for annotating plastome sequences. Appl Plant Sci 3: . doi: 10.3732/apps.1500026

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:868–877

Hunter JD (2007) Matplotlib: A 2D Graphics Environment. Comput Sci Eng 9:90–95

Joshi NA FJN (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.210). Available at https://github.com/najoshi/sickle

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359

Li H, Handsaker B, Wysoker A, et al. (2009) The Sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Linder CR, Goertzen LR, Heuvel BV, et al. (2000) The complete external transcribed spacer of 18s-26s rdna: amplification and phylogenetic utility at low taxonomic levels in Asteraceae and closely allied families. Mol Phylogenet Evol 14:285–303

Malé P-JG, Bardon L, Besnard G, et al (2014) Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. Mol Ecol Resour 14:966–975

McKenna A, Hanna M, Banks E, et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

McKinney W (2010) Data structures for statistical computing in python. Proc. of the 9th Python in Science Conference

Michael TP (2014) Plant genome size variation: bloating and purging DNA. Brief Funct Genomics 13:308–317

Norquist C (1991) Recovery Plan: Mohr’s Barbara's Buttons ('*Marshallia mohrii*') Beadle and FE Boynton. U.S. Fish and Wildlife Service, Jackson Mississippi

Pedregosa F, Varoquaux G, Gramfort A, et al. (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Pinheiro J, Bates D, DebRoy S, et al. (2017) Package “nlme.” Linear and nonlinear mixed effects models 3–1

Rambaut A (2009) FigTree. Tree figure drawing tool version 1.3. 1. Institute of Evolutionary biology, University of Edinburgh

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277

Ruby JG, Bellare P, Derisi JL (2013) PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 3:865–880

Rutherford K, Parkhill J, Crook J, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945

Smit A, Hubley R, Green P (2017) 1996--2010. RepeatMasker Open-3.0

Soltis DE, Visger CJ, Marchant DB, Soltis PS (2016) Polyploidy: Pitfalls and paths to a paradigm. Am J Bot 103:1146–1166

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Staton SE, Burke JM (2015) Transposome: a toolkit for annotation of transposable element families from unassembled sequence reads. Bioinformatics 31:1827–1829

Straub SCK, Parks M, Weitemier K, et al (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am J Bot 99:349–364

Swofford DL (2003) PAUP*: phylogenetic analysis using parsimony, version 4.0 b10

U.S. Fish And Wildlife (1988) Endangered and threatened wildlife and plants; determination of *Marshallia mohrii* (Mohr’s Barbara’s buttons) to be a threatened species. Fed Regist 53:4698–34701

Van der Auwera GA, Carneiro MO, Hartl C, et al. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11.10.1–33

Van Rossum G (2014) Python 3.4.3. https://www.python.org/downloads/release/python-343/

Vargas OM, Ortiz EM, Simpson BB (2017) Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: *Diplostephium*). New Phytol 214:1736–1750

Wang K, Hong W, Jiao H, Zhao H (2017) Transcriptome sequencing and phylogenetic analysis of four species of luminescent beetles. Sci Rep 7:1814

Watson LE, Estes JR (1987) Chromosomal evolution of *Marshallia* (Asteraceae). Am J Bot 74:764

Watson LE, Estes JR (1990) Biosystematic and Phenetic Analysis of *Marshallia* (Asteraceae). Syst Bot 15:403–414

Weakley AS, Poindexter DB (2012) A new species of *Marshallia* (Asteraceae, Helenieae, Marshalliinae) from mafic woodlands and barrens of North Carolina and Virginia

Wickett NJ, Mirarab S, Nguyen N, et al (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A 111:E4859–68

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255

Xu W, Wongsa A, Lee J, et al (2011) RNA2DMap: A visual exploration tool of the information in RNA's higher-order structure. Proceedings 613–617

Zhang C, Sayyari E, Mirarab S (2017) ASTRAL-III: increased scalability and impacts of contracting low support branches. In: Meidanis J, Nakhleh L (eds) Comparative Genomics. Springer, Cham, pp 53–75

Table 1 Sequence cleaning information DNA. Organized by species and sample name.

Species	sample name	method	total fragments	passed fragments	per. filter	filter fragments
Marshallia caespitosa	M26_9	afterQC	40999110	38780777	5.4	2218333
	M28_7	afterQC	35545885	34089789	4.1	1456096
Marshallia graminifolia	M1_16	afterQC	15885380	15801948	0.5	83432
	M5_1	afterQC	44126317	42693915	3.2	1432402
	M39	afterQC	20324114	19012571	13.0	1311543
Marshallia grandiflora	M30_2	afterQC	31263991	30709818	1.8	554173
	M17	afterQC	43766503	41233309	5.8	2533194
Marshallia legrandii	M31_5	afterQC	30494968	29580274	3.0	914694
Marshallia mohrii	M21_1	afterQC	35858590	34854640	2.8	1003950
	M24_1	afterQC	34811147	33561826	3.6	1249321
Marshallia obovata	M32_2	afterQC	32139693	30875050	3.9	1264643
	M3_1	afterQC	14472738	14365053	0.7	107685
	M4_1	afterQC	30111244	28907600	4.0	1203644
Marshallia ramosa	M19_2	afterQC	33195043	32237473	2.9	957570
	M38	afterQC	18955698	17694819	13.4	1260879
Marshallia trinervia	*M33_6	afterQC	52213583	25712687	49.0	25559165
	M2_3	afterQC	19562142	19114627	2.3	447515

*AfterQC failed to clean the reads from 1 of the 3 runs which reduced the number of reads, however, given the total number of reads, there was enough for the genome skimming process.

Table 2 Sequence cleaning information for cDNA

Species	sample	sequencing	method	total	passed	per. filter	filtered
Marshallia caespitosa	SL148333	illumina	afterqc	25590756	21373533	16.5	4217223
	SL148330	illumina	afterqc	58683007	49372966	15.9	9310041
	SL148332	illumina	afterqc	47945454	41467482	13.5	6477972
	SL148331	illumina	afterqc	42149008	36412176	13.6	5736832
	SL148327	illumina	afterqc	39842513	31404514	21.2	8437999
	SL148328	illumina	afterqc	46553274	40385240	13.2	6168034
	SL148329	illumina	afterqc	48866310	39964079	18.2	8902231
	SL148334	illumina	afterqc	47567999	37896649	20.3	9671350
Marshallia mohrii	M20r	illumina	afterqc	26286499	21169085	19.5	5117414
	M21r	illumina	afterqc	51223374	39755591	22.3	11467783
Marshallia grandiflora	M29r	illumina	afterqc	25844057	23282510	9.9	2561547
Marshallia trinervia	M2_9r	illumina	afterqc	19011413	13227391	30.4	5784022
Marshallia ramosa	M38r	illumina	afterqc	33463481	31019283	7.3	2444198
Marshallia graminifolia	M39r	illumina	afterqc	43933352	40622949	7.5	3310403
Marshallia obovata	M3_9r	illumina	afterqc	14982599	8838696	41.0	6143903
Helenium autumnale	SRR1141042	illumina	afterqc	12606640	12602396	0.0	4244
Guizotia abyssinica	SRR5237269	illumina	afterqc	42316316	42073316	0.6	243000
Glebionis segetum	SRR797217	illumina	afterqc	13015242	11612524	10.8	1402718
Guizotia scabra	SRR797218/ SRR797219	454	cutadapt	1267082	1267082	0	0
*Gaillardia aristata**	SRR5237184	illumina	trim	40820201	NA	NA	NA
*Inula racemosa**	SRR5237205	illumina	trim	47085168	NA	NA	NA

*single end sequences were cleaned within assembly, or pipeline using builtin cleaning trimmomatic or trim_galore, no stand alone filtered fastqs were produced.

Table 3 Trinity and Trinotate summaries

Species	sample	unique genes	unique transcripts	annotations
Gaillardia aristata	SRR5237184	155037	274583	278561
Gaillardia pulchella	SRR5237268	130650	248626	250889
Glebionis segetum	SRR797217	40967	116039	116425
Guizotia abyssinica	SRR5237269	81433	215740	219903
Helenium autumnale	SRR1141042	51443	85534	87043
Inula ensifolia	SRR5237206	172918	308927	329752
Inula racemose	SRR5237205	144603	254798	285926
Marshallia caespitosa	M10 SL148332	98018	250301	265013
Marshallia mohrii	M20r	83156	186139	192518
	M21r	102665	246496	256025
Marshallia trinervia	M2_9r	46220	66005	66949
Marshallia grandiflora	M29r	50940	89355	91038
Marshallia ramosa	M38r	78604	163167	167621
Marshallia obovata	M3_9r	52602	89798	92535
Marshallia graminifolia	M39r	59233	101300	103229

* a representative sample was chosen for annotation, but all M10 assemblies were pooled, for cd-hit and OrthoFinder analysis

Table 4 Orthogroups by the number of genes they are predicted to contain produced by OrthoFinder

avg. no. genes	no. ortho	per. ortho	no. genes	per. genes
less than 1	16951	60.5	81396	16.8
1	6633	23.7	138901	28.7
2	2339	8.3	83688	17.3
3	964	3.4	49147	10.2
4	461	1.6	30414	6.3
5	233	0.8	18986	3.9
6	117	0.4	11242	2.3
7	88	0.3	9822	2
8	52	0.2	6549	1.4
9	29	0.1	4114	0.9
10	35	0.1	5508	1.1
11-15	59	0.2	11317	2.3
16-20	27	0.1	7348	1.5
21-50	26	0.1	11775	2.4
51-100	6	0	6509	1.3
101-150	2	0	4047	0.8
151-200	1	0	2509	0.5

Table 5 Percentage of unfiltered reads per accession and per species that mapped to *Del/Tekay Marobo* (NCBI Accession: KX396599.1)

Species	Percent per species	Percent per sample	total raw reads	sample	reads mapped
Marshallia caespitosa	0.0260	0.0284	81998220	M26	23301
		0.0235	71091770	M28	16697
Marshallia graminifolia	0.0850	0.1002	31770760	M1	31833
		0.0962	59967234	M39	57702
		0.0586	88252634	M5	51690
Marshallia grandiflora	0.0936	0.0964	87533006	M17	84372
		0.0907	62527982	M30	56726
Marshallia legrandii	0.0621	0.0621	60989936	M31	37891
Marshallia mohrii	0.0514	0.0508	71717180	M21	36408
		0.0521	69622294	M24	36275
Marshallia obovata	0.0546	0.0516	28945476	M3	14945
		0.0547	64279386	M32	35146
		0.0575	60222488	M4	34610
Marshallia ramosa	0.0545	0.0536	66390086	M19	35593
		0.0553	48878510	M38	27045
Marshallia trinervia	0.0172	0.0198	39124284	M2	7732
		0.0147	79793124	M33	11706

Table 6 Superfamily abundance by subsample, with the minimum cluster size for detection, and number of reads used in analysis

species	sample	Subsample	thousands of reads used	minimum cluster size	total	Copia	Gypsy	Harbinger	Helitron	L1	MuDR	hAT	unclassified
caes.	M26	1	400	100	0.695	0.376	0.274	0.009	0.027	0.005	0.003	0.001	0.000
		2	400	100	0.694	0.401	0.245	0.007	0.029	0.005	0.004	0.001	0.000
		3	400	100	0.694	0.365	0.287	0.005	0.028	0.005	0.003	0.001	0.000
	M28	1	400	100	0.683	0.381	0.250	0.008	0.031	0.005	0.005	0.002	0.000
		2	400	100	0.684	0.389	0.248	0.011	0.029	0.004	0.003	0.000	0.000
		3	400	100	0.684	0.381	0.252	0.010	0.032	0.004	0.005	0.000	0.000
grami.	M1	1	50	25	0.570	0.253	0.271	0.023	0.018	0.001	0.001	0.004	0.000
		2	400	100	0.736	0.374	0.302	0.026	0.026	0.005	0.003	0.000	0.000
		3	400	100	0.736	0.383	0.294	0.027	0.026	0.003	0.003	0.000	0.000
	M39	1	400	100	0.731	0.336	0.349	0.025	0.014	0.004	0.002	0.000	0.000
		2	400	100	0.731	0.337	0.340	0.027	0.016	0.004	0.002	0.005	0.000
		3	400	100	0.732	0.325	0.355	0.025	0.016	0.004	0.002	0.004	0.000
	M5	1	400	100	0.696	0.368	0.262	0.030	0.029	0.003	0.005	0.000	0.000
		2	400	100	0.696	0.366	0.266	0.030	0.026	0.005	0.004	0.000	0.000
		3	400	100	0.694	0.368	0.257	0.036	0.022	0.003	0.004	0.004	0.000
grandi.	M17	1	400	100	0.682	0.310	0.310	0.014	0.037	0.008	0.004	0.000	0.000
		2	400	100	0.682	0.331	0.290	0.013	0.034	0.010	0.004	0.001	0.000
		3	400	100	0.682	0.312	0.299	0.017	0.038	0.010	0.006	0.000	0.000
	M30	1	400	100	0.716	0.320	0.340	0.013	0.032	0.009	0.003	0.000	0.000
		2	400	100	0.715	0.319	0.339	0.014	0.029	0.011	0.004	0.000	0.000
		3	400	100	0.714	0.296	0.362	0.012	0.030	0.011	0.003	0.000	0.000
leg.	M31	1	400	100	0.684	0.317	0.297	0.018	0.036	0.012	0.004	0.000	0.000
		2	400	100	0.683	0.317	0.297	0.016	0.039	0.011	0.003	0.001	0.000
		3	400	100	0.683	0.315	0.299	0.014	0.039	0.012	0.005	0.000	0.000
mohr.	M21	1	400	100	0.682	0.350	0.270	0.015	0.033	0.010	0.003	0.002	0.000
		2	400	100	0.680	0.376	0.255	0.006	0.028	0.009	0.003	0.003	0.000
		3	400	100	0.682	0.391	0.238	0.009	0.033	0.008	0.003	0.000	0.000
	M24	1	400	100	0.695	0.393	0.247	0.011	0.032	0.008	0.003	0.000	0.000
		2	400	100	0.696	0.384	0.260	0.008	0.030	0.008	0.004	0.003	0.000
		3	400	100	0.697	0.377	0.268	0.010	0.030	0.007	0.003	0.002	0.000
obo.	M32	1	50	25	0.574	0.252	0.259	0.016	0.033	0.010	0.002	0.001	0.000
		2	50	25	0.572	0.265	0.264	0.012	0.025	0.006	0.001	0.000	0.000
		3	50	25	0.572	0.238	0.274	0.019	0.038	0.002	0.002	0.000	0.000

	M3	1	50	25	0.605	0.223	0.322	0.016	0.027	0.013	0.002	0.002	0.000
		2	50	25	0.611	0.252	0.305	0.008	0.036	0.003	0.006	0.001	0.000
		3	50	25	0.608	0.263	0.303	0.012	0.020	0.004	0.003	0.002	0.000
	M4	1	400	100	0.727	0.355	0.304	0.018	0.039	0.008	0.003	0.000	0.000
		2	400	100	0.727	0.336	0.323	0.019	0.036	0.008	0.004	0.002	0.000
		3	400	100	0.724	0.332	0.331	0.016	0.035	0.008	0.002	0.000	0.000
ram.	M19	1	400	100	0.716	0.385	0.277	0.014	0.027	0.008	0.004	0.000	0.000
		2	400	100	0.715	0.412	0.248	0.015	0.026	0.005	0.004	0.004	0.000
		3	400	100	0.715	0.403	0.265	0.012	0.024	0.006	0.005	0.000	0.000
	M38	1	400	100	0.720	0.413	0.254	0.014	0.030	0.006	0.004	0.000	0.000
		2	400	100	0.719	0.395	0.267	0.014	0.033	0.006	0.003	0.000	0.000
		3	400	100	0.720	0.404	0.266	0.011	0.032	0.005	0.002	0.000	0.000
trin.	M2	1	400	100	0.693	0.385	0.225	0.019	0.047	0.013	0.003	0.000	0.000
		2	400	100	0.692	0.384	0.226	0.018	0.045	0.015	0.004	0.001	0.000
		3	400	100	0.692	0.385	0.230	0.017	0.040	0.015	0.004	0.000	0.000
	M33	1	400	100	0.717	0.387	0.263	0.012	0.033	0.017	0.004	0.000	0.000
		2	400	100	0.717	0.372	0.275	0.019	0.033	0.012	0.005	0.001	0.000
		3	400	100	0.716	0.370	0.270	0.022	0.034	0.016	0.004	0.000	0.000

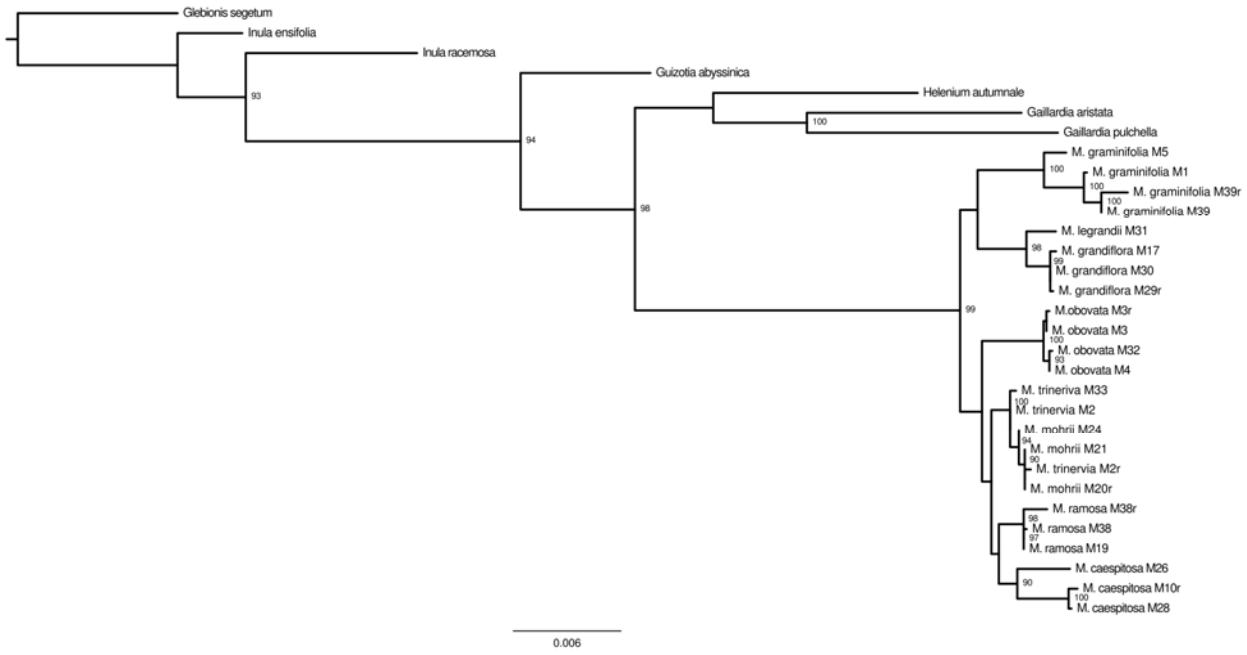


Figure 1 Rooted rDNA tree containing rDNA derived from transcriptomic and genomic sources. Bootstrap values greater than or equal to 90 are shown.

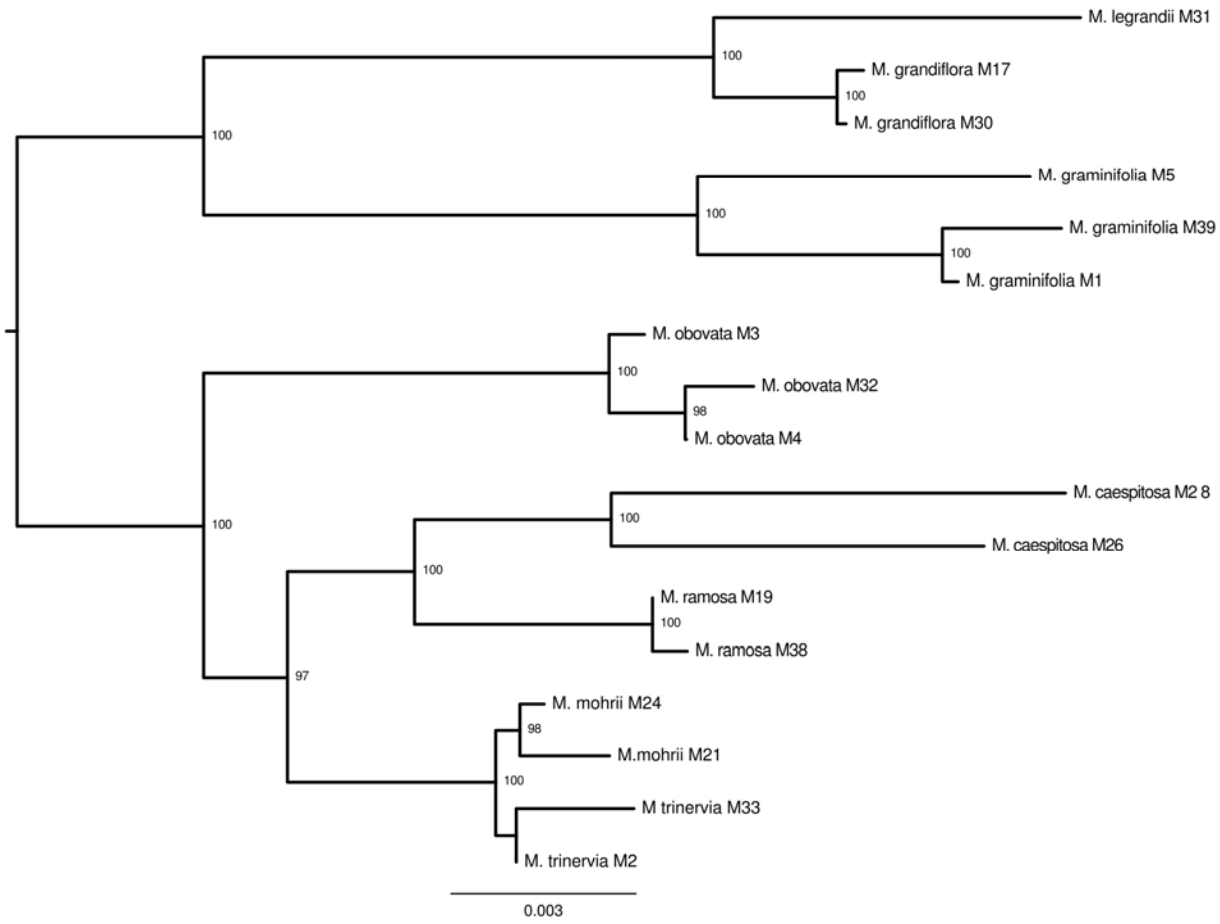


Figure 2. Ribosomal DNA tree derived from the alignment of genomic rDNA assembly. Rooting determined by previous analysis of combined rDNA tree. Bootstrap values greater than or equal to 90 are shown.

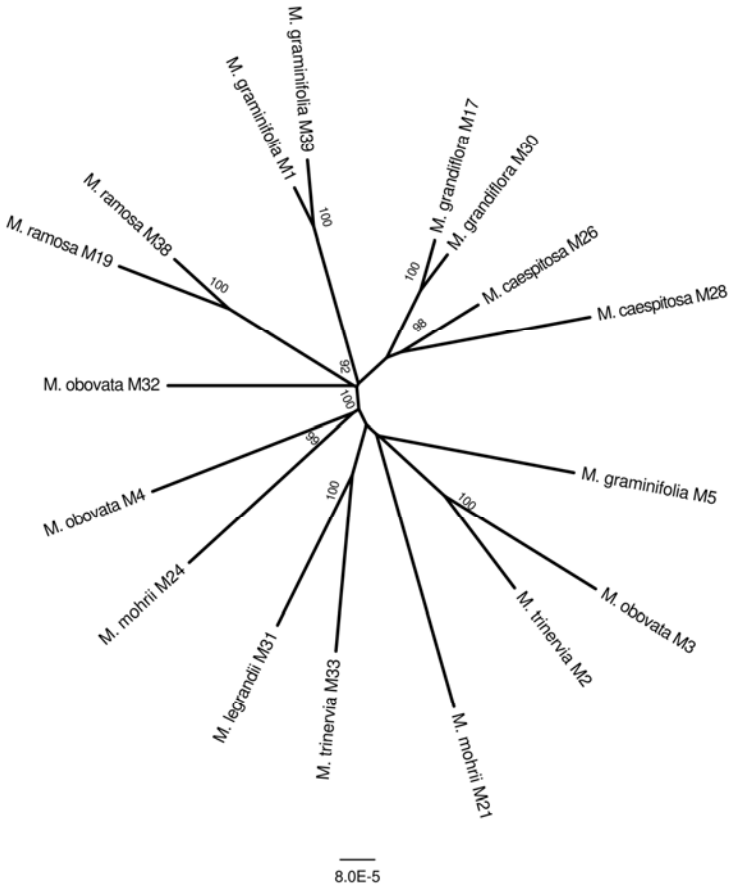


Figure 3 Unrooted maximum likelihood tree of whole *Marshallia* plastid sequence produced with RAxML with node support determined by 1000 bootstraps and values greater than or equal to 90 are shown

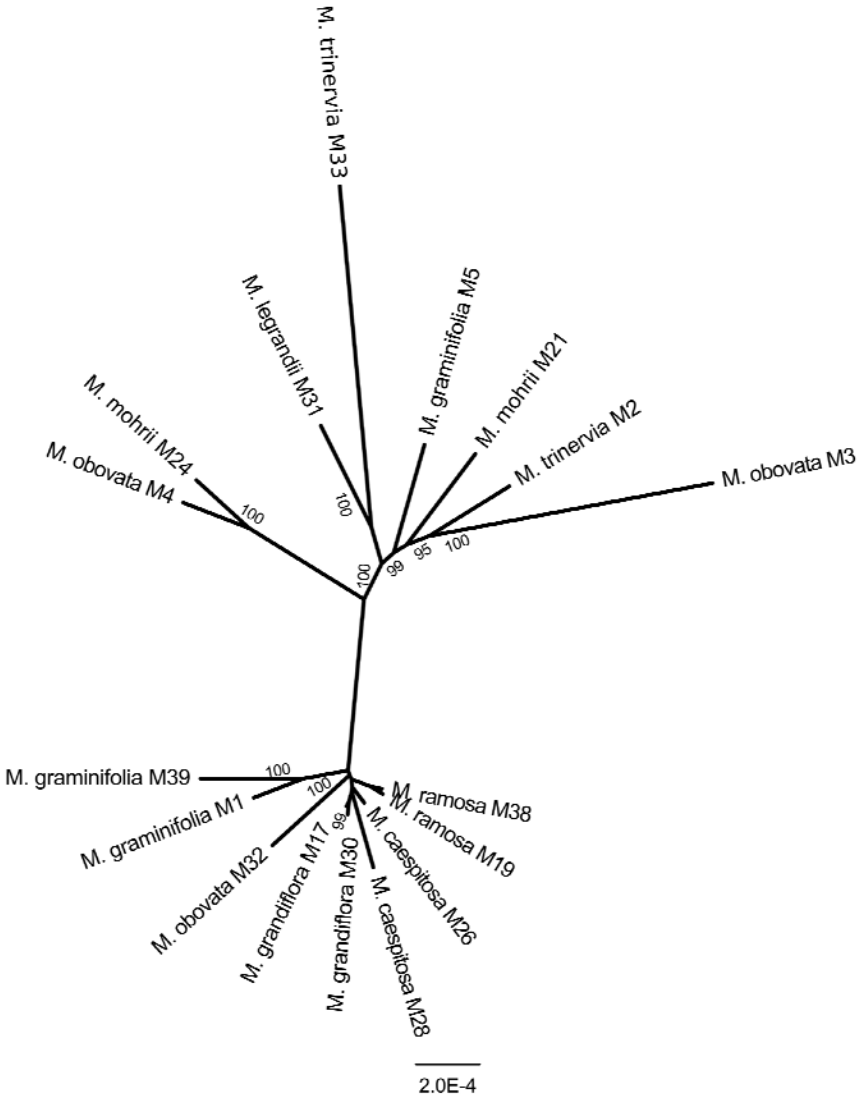


Figure 4. Unrooted maximum likelihood tree of *Marshallia* mitochondrial sequence produced with RAxML with node support determined by 1000 bootstraps and values greater than or equal to 90 are shown.

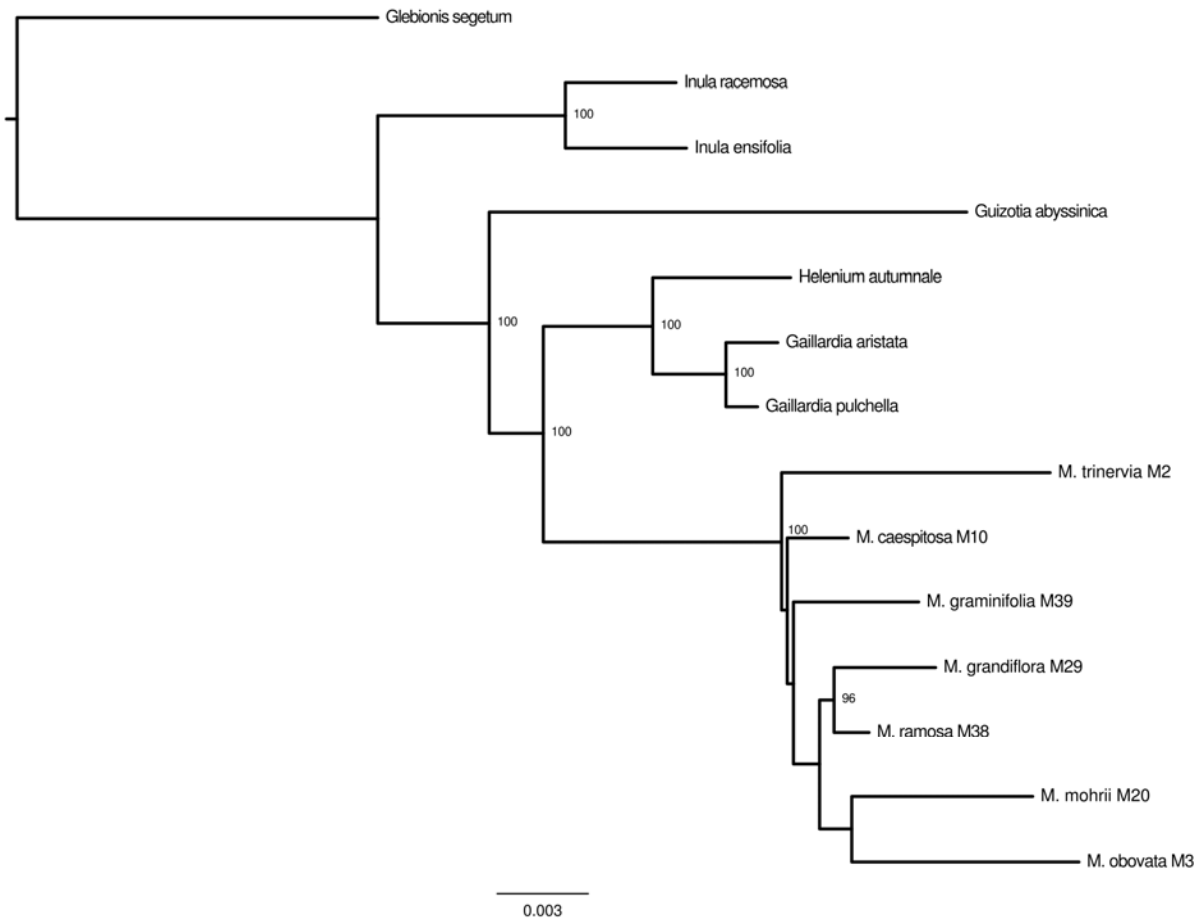


Figure 5 Transcriptomic coding sequences for plastid and mitochondrial genomes

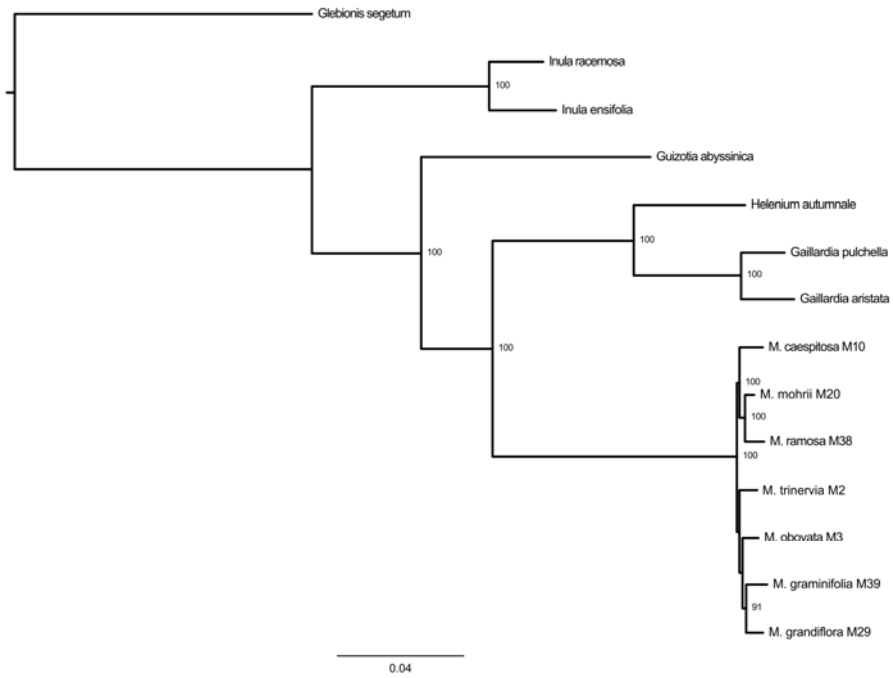


Figure 6. Rooted RAxML tree with *Marshallia* and outgroups. RAxML run with GTR + Gamma model with node supported determined by 1000 bootstraps

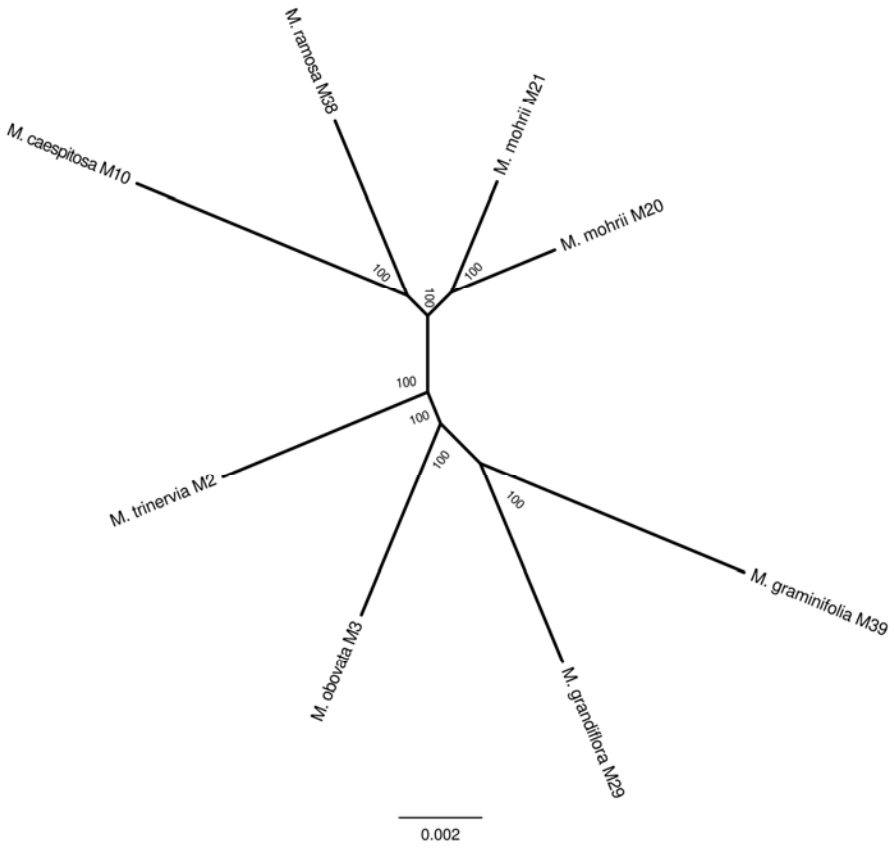


Figure 7 Unrooted RAxML tree for single copy trinity assembled transcripts within *Marshallia.* This tree is representative of both the relaxed and strict cleaning approaches used prior to RAxML run with GTR + Gamma model with node supported determined by 1000 bootstraps.

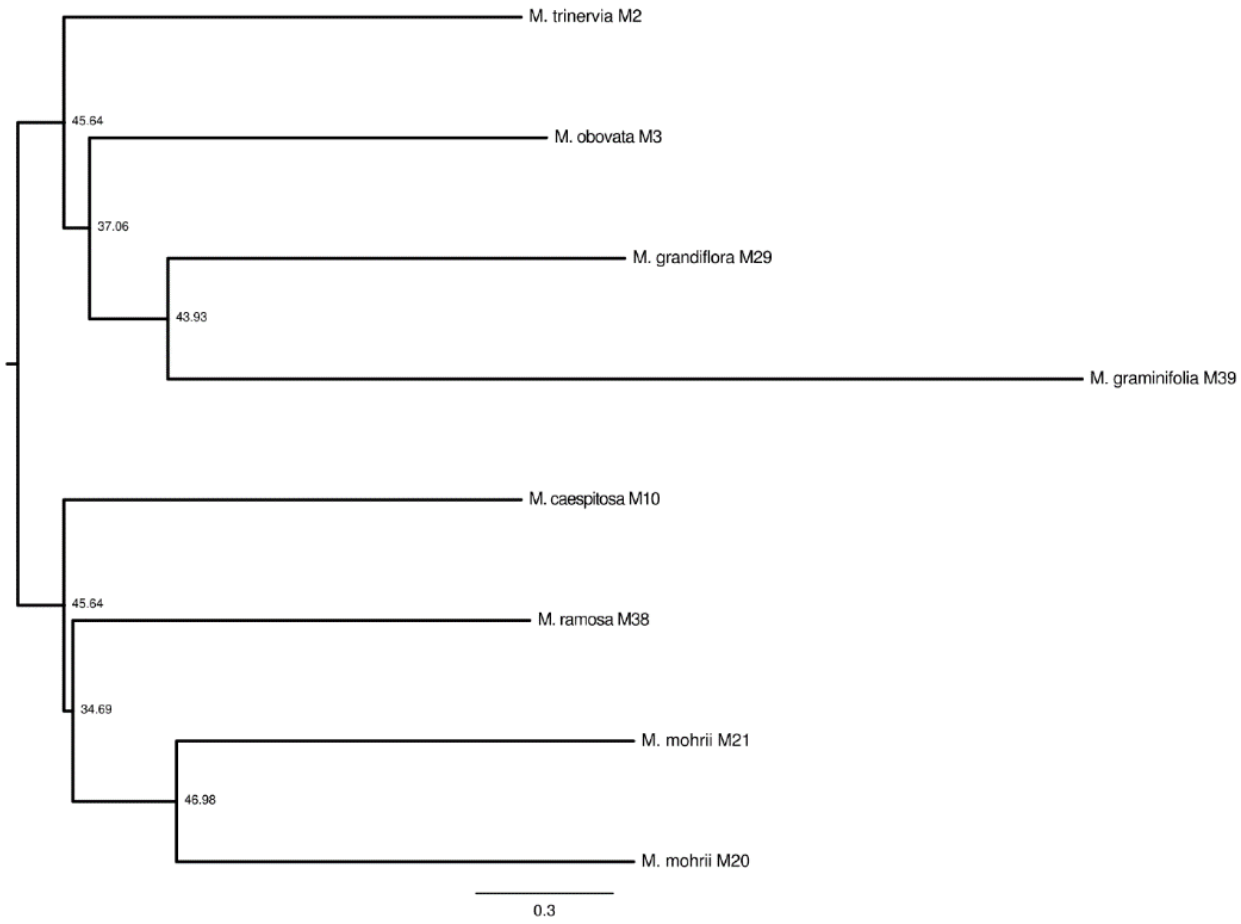


Figure 8 Astral tree nodes labeled with local support for each node from all the gene trees together

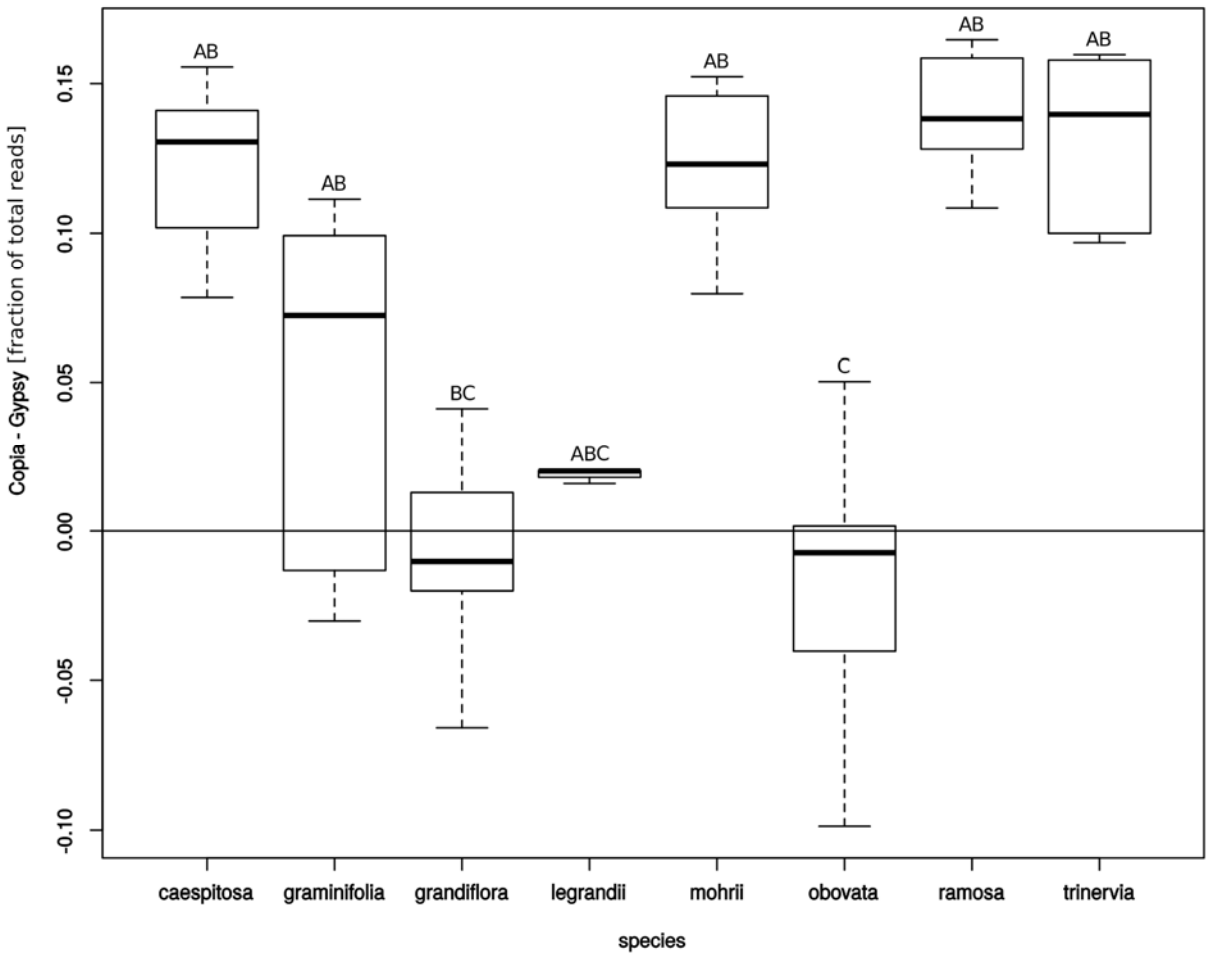


Figure 9 Grouping of differences between the fraction of Copia and Gypsy superfamilies among all species of *Marshallia.* Grouping was determined Tukey's HSD applied to a linear mixed model, which controlled for repeated measures within each sample.

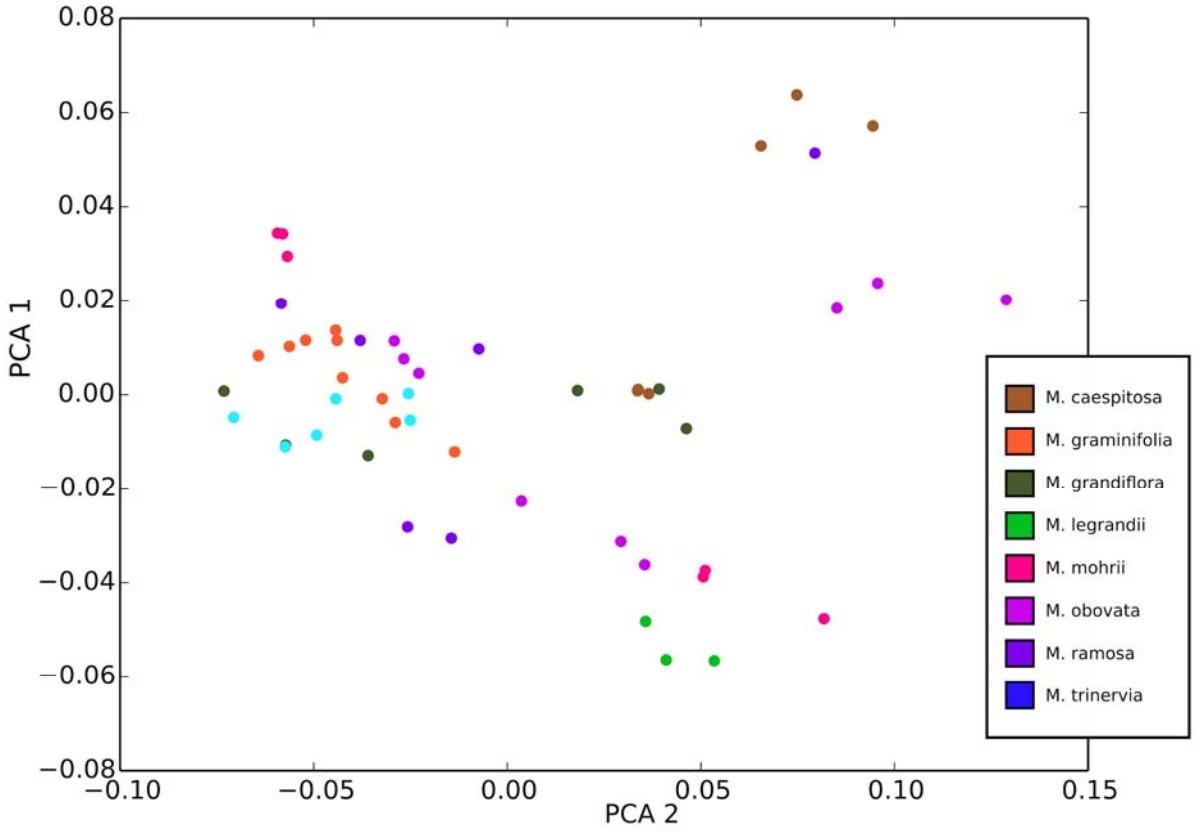


Figure 10 PCA of repetitive element superfamily abundance