The Thesis Committee for Aldanah Ayidh A. Alqahtani
Certifies that this is the approved version of the following Thesis:

Genomic resources of *Euphorbia schimperi*: Evolutionary and medicinal

APPROVED BY
SUPERVISING COMMITTEE:

Robert K Jansen, Supervisor

Edward C Theriot

Genomic resources of *Euphorbia schimperi*: Evolutionary and medicinal

by

Aldanah Ayidh A. Alqahtani

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

December 2018

Dedication

To my beloved grandmother Aldanah (god bless her soul), grandfather Masud, My parents, My aunt Monura, My brother Masud, My professor Neven, My friends; Haifa Alqarani, Dalia Halwani, Samah Alsalkadi and her family.

Your love, kindness, encouragement, and positive influence are appreciated

.

Acknowledgements

Above all, I give thanks to Allah for his countless blessings.

This work would not have been conceivable without contributions from several individuals and organizations. Their assistance and support had been substantially significant throughout my graduate studies at the University of Texas at Austin.

First, I would like to express my sincerest gratitude to my supervisor Dr. Jansen K. Robert for his invaluable guidance, support, and encouragement throughout my study. It is a must for me to emphasize the great level of organization and the continuous weekly advising session he provided, which helped me tackle various challenges in my research. I am privileged to have had an opportunity to work under his supervision. Through him I learned and grew a great deal as a student, an individual, and a professional. I am honored to have been a student of Dr. Jansen and will always be grateful for the professional advice he has given me. In addition, my deepest appreciation goes to my second reader Dr. Theriot C. Edward for reading my work, and contributing insightful comments.

Furthermore, a great deal of appreciation goes to Dr. Ruhlman Tracy for her help and guidance throughout the DNA&RNA extraction experiments. Many thanks go to Benjamin Goetz and Dhivya Arasappan from the Center for computational Biology and Bioinformatics (CCBB) at the university of Texas at Austin for their help in my transcriptome analysis.

I also appreciate my fellow graduate students, Bikash Shrestha, Chaehee Lee, Insue Choi, Deise Goncalves, and Edgardo Valencia, who shared their expertise in the lab with me. Additionally, the library at the University of Texas at Austin is appreciated for providing access to resources and journal papers reviewed during my research.

Finally, I would like to acknowledge my academic sponsors at Prince Sattam bin Abdul-Aziz for providing me with the opportunity to further my education and pursue a master's degree.

Abstract

Genomic resources of *Euphorbia schimperi*: Evolutionary and medicinal

Aldanah Ayidh A. Alqahtani, M.A

The University of Texas at Austin, 2018

Supervisor: Robert K. Jansen

Euphorbia schimperi is a toxic plant within the Euphorbiaceae with medicinal applications. The species is known to produce secondary compounds that are effective against breast and brain cancer but very little is currently known about the genomics of *E. schimperi*. In order to enhance the understanding of phylogenetic relationships with other plants within the same family, the plastid genome (plastome) was sequenced with the aim of understanding its evolution. The results revealed *E. schimperi* plastome is 159,463 bp in size and it includes a pair of inverted repeats (IR) of 26,629 bp, which are separated by a small single copy (SSC) region of 17,305 bp and a large single copy (LSC) region of 88,900 bp. *Euphorbia schimperi* plastome has a GC content of 35.6. The total number of protein coding genes is 84 while the total number of genes is 118. In the IR, 19 genes are duplicated and there are 30 and 4 tRNA and rRNA genes, respectively. However, in *E. schimperi*, the *infA*, *rps16* and *rpl32* genes are non-functional. The medicinal plant, *E. schimperi*, accumulates important secondary compounds, such as triterpenoids like cycloart-23-en-3β, 25-diol, α-amyrin, and cycloart-25-en-3β 24-diol as well as steroid glycosides, flavonoids and phenylpropanoids, which are known to have brain and breast anti-cancer properties. In this study, transcriptome sequencing of *E. schimperi* was conducted to identify the

enzymes (genes) that are involved in biosynthesis of these compounds. About 80,916,952 million paired-end sequence reads of 150 bp were generated from RNA isolated from fresh leaf tissue of *E. schimperi* using IIlumina platform. Gene family trees were constructed to identify homologs of six pathway genes (LUP2, AT4G34050 (CCOAOMT1), FNSI, UGT80A2, RcCAS, and FLS1) to examine their evolutionary relationship. Sequences of all six genes were identified in the transcriptome of *E. schimperi* and some of them have more copies then the other members of Malpighiales as well as Fabales. The transcriptome study provides valuable information for understanding specialized plant metabolism and genomic resources for increasing the production of plant- derived pharmaceuticals.

Table of Contents

List of Tables

List of Figures

Chapter1: Complete chloroplast genome sequence of *Euphorbia schimperi*

1.1. Introduction

The Euphorbiaceae has approximately 7,500 species organized into 300 genera, 37 tribes, and three subfamilies (Bramwell & Bramwell, 2001). It is one of the largest families of angiosperms and contains approximately ten species that exhibit promising anticancer activity (Bhanot et al., 2011). *Euphorbia* contains over 2000 species, making it one of the largest genera of flowering plants (Buddensiek, 2005). The genus has unique aesthetic floral features and latex-like sap in their stems. The primary purpose of this sap is to protect the plants from herbivores and it has been used to treat cancer (Jade, 2012; Pharm et. al, 2017). *Euphorbia schimperi C.Presl* grows mainly as a succulent shrub in rocky environments of open savannahs and is a perennial plant reaching heights of 1 to 1.7 meters. The distribution of *E. schimperi* is the southern part of the Arabian Peninsula in Saudi Arabia, Yemen and Oman as well as in Northeast Africa.

Only six of the species in the Euphorbiaceae have known plastome sequences, including *Jatropha curcas, Hevea brasiliensis*, *Ricinus communis*, *Manihot esculenta, Euphorbia esula*, and *Vernicia fordii*. These species are economically important and have been shown to have some medicinal activities due to the presence of isoprenoids (Ahmed et al., 2016).

One characteristic that makes angiosperm plastomes distinctive is the presence of a quadripartite structure, which is composed of a small single copy region, a large

single copy region and a pair of inverted repeats (Li et al., 2017; Rivarola et al., 2011; Tangphatsornruang et al., 2011). The angiosperm plastome is also known for its high level of conservation of gene order and gene content (Ruhlman and Jansen 2014). Compared to the nuclear genome, the rate of nucleotide substitutions is relatively lower in plastomes (Jansen and Ruhlman 2012). The stable structure and predominantly uniparental inheritance make the plastome very valuable for evolutionary studies. In certain genes within specific lineages, a rapid rate of evolution has been observed (Li et al., 2017). Several genes, including *ycf1, matK* and *rbcL* are used as essential DNA barcodes to identify plants (Cauz-Santos et al., 2017; Li et al., 2017). Due to the above characteristics, plastomes are good models for molecular evolution investigations.

The expected result of chloroplast sequencing analysis should be a genome size between 150,000 bp -164,000 bp based on information from other Euphorbiaceae species (Tangphatsornruang et al., 2011). The number of genes in Euphorbiaceae plastomes ranges from 118-135, and proteins involved in gene expression are encoded by approximately 84 genes. Other genes code for four rRNAs and 30 tRNAs. Moreover, the seven Euphorbiaceae plastomes have four distinct regions, including a pair of inverted repeats (IRs, 26-28kb), which are separated by a small single copy (SSC,17-19kb) region and a large single copy(LSC,89-92kb) region (Rivarola et al., 2011; Asif et al., 2010; Cauz-Santos et al., 2017; Tangphatsornruang et al., 2011).

Using next generation sequencing technologies followed by *de novo* assembly, the present study reconstructed the whole plastome of *Euphorbia schimperi.* The

objective of the study is to gain insight into the molecular evolution of the plastome through characterization of the complete plastome and comparative genomics with other members of Euphorbiaceae. In addition, phylogenetic relationships with closely related species of *Euphorbia schimperi* were investigated.

1.2. Methods

1.2.1. Plant material and DNA isolation

Fresh leaf tissues of *Euphorbia schimperi* were collected from the Arid Land Greenhouses https://aridlandswholesale.com/ in Tucson, Arizona. A voucher specimen of *E. schimperi* was deposited in TEX herbarium. Whole genomic DNA was extracted from 0.2 g of the leaves based on the method proposed by Doyle and Doyle (1987).

1.2.2. Organelle Genome sequencing

One DNA sample with a concentration of 100ng/µl and volume of 42 µl was submitted for sequencing to the Genome Sequencing Analysis Facility (GSAF, University of Texas at Austin) for paired end sequencing (2 x150 bp) on the Illumina HiSeq 4000 platform. The minimum number of reads expected was 20 million reads while the targeted reads were 30 million reads.

1.2.3. Genome Assembly

The paired end Illumina reads totaled 65,948,636 bp and were assembled *de novo* via Velvet with multiple K-mer values as suggested by Zerbino & Birney (2008). In Geneious Version 10.0.6, the initial plastid reads were assembled into two long contigs that included the entire plastome. Gaps between *ycf1* and *ndhF* were filled by

overlapping contigs. The other gap between *atpH* and *atpF* and ambiguous nucleotides was filled by mapping against the raw reads using Bowtie 2 version 2.3.4 (Langmead & Salzberg, 2012). Mapping the raw reads to the *E. schimperi* plastome indicated that the average coverage was 1465 reads.

1.2.4. Genome Annotation

Annotation of the plastome conducted using multiple software platforms. Geneious vers. 10.0.6 checked for the start and stop codons for every gene compared to tobacco (*Nicotiana tabacum*) and other species of Euphorbiaceae, including *Jatropha curcas, Euphorbia esula, Manihot esculenta, Ricinus communis,* and *Hevea brasiliensis.* Dual Organellar Genome Annotator (DOGMA) was utilized to identify coding sequences with default settings (Wyman et al., 2004). BLAST homology searches were used to search for rRNAs, tRNAs and protein coding genes followed by manual correction of any start, stop codons as well as intron position. The tRNAscan-SE online search server was used to confirm all identified tRNA genes (Low & Eddy, 1997; Low & Chan, 2016). Based on the loss of portions of the sequence or presence of internal stop codons, pseudogenes were identified. A genome map was drawn using OGDRAW (Lohse et al., 2013).

1.2.5. Comparative Analysis of Plastome

The complete plastome of *E. schimperi* was compared to other closely related species from the same family (*Jatropha curcas, Euphorbia esula, Ricinus communis, Hevea brasiliensis, Vernicia fordii* and *Euphorbia esaul*) in overall size, in LSC, SSC, and IR length, GC content %, and in total number of genes, genes duplicated in IR,

protein coding genes, tRNA and rRNA genes. The border position of LSC, SSC and IR regions among sequences of eight species (*J. curcas, E. esula, R. communis, V. fordii, M. esculenta*, *H. brasiliensis* and *N. tabacum*) including *Euphorbia schimperi* was compared. The Progressive Mauve algorithm (Darling et al., 2004) in Geneious Version 10.0.6 was used to examine conserved regions and gene order among seven species of Euphorbiaceae compared to *Nicotiana tabacum*.

1.2.6. Phylogenetic analysis

A phylogenetic analysis of Euphorbiaceae was conducted using maximum likelihood (ML) tree of plastome sequences using RAxML v8. with the 'GAMMA GTR' model under rapid bootstrapping algorithm with 1000 bootstrap replicates in Geneious Version 10.0.6. A majority rules consensus tree was generated with 50% bootstrap value support threshold. Plastome sequences were aligned using MAFFT alignment V7.3 88 (Katoh and Standley, 2013) followed by manual adjustment in Geneious Version 10.0.6 for *Euphorbia schimperi*, *Jatropha curcas, Ricinus communis, Vernicia fordii, Euphorbia esaul*, *Manihot esculenta, Hevea brasiliensis* and *Nicotiana tabacum*as outgroup).

1.3. Results and Discussion

1.3.1. General features of *Euphorbia schimperi Plastome*

The *Euphorbia schimperi* plastome has a length of 159,463 base pairs (bp) with a pair of inverted repeats (IR) of 26,629 bp, which separate the LSC (88,900 bp) and SSC (17305 bp) (**Figure 1.1** and **Table 1.1**). The genome encodes a total of 118 different genes including 4 *rRNA* genes, 30 *tRNA* genes and 84 protein-coding genes (**Table 1.1**).

The complete plastid genome of *Euphorbia schimperi*, was sequenced and revealed the losses of two genes, translation initiation factor 1 (*infA)*, the ribosomal protein L32 (*rpl32*) *and pseudogene ribosomal protein S16 (rps16).*

1.3.2. Pseudogenization of *rps16* Gene and the two Missing Genes *infA* and *rpl32*

In photosynthetic plants, gene loss from the plastome rarely takes place except in instances where the gene has been functionally transferred to the nucleus (Magee, et al., 2010). The number of genes and their order are generally highly conserved in plastomes (Roy, et al., 2010). *In this study, Euphorbia schimperi has lost two genes (infA*, rpl*32) and has one pseudogene (rps16). In the previous studies rps16* was lost from the plastomes of *Medicago* (Saski et al., 2005), *Pinus* (Wakasugi et al., 1994 (Ueda et al., 2007), *Jatropha curcas* (Asif et al., 2009) *, Populus trichocarpa* (Steane, 2005) and *Arabis stellari* (Raman et al., 2017). *In addition, it has been reported that* the most common gene loss involves *infA*, with at least 11 independent losses in angiosperms (Jansen et al., 2007).The *infA* gene was found to be a pseudogene or entirely missing from *Dianthus superbus* and *Lychnis wilfordii* of the Caryophyllales order, as well as Brassicales, Cucurbitales, Fabales, Malpighiales (Euphorbiaceae; *Jatropha curcas*, *Manihot esculenta, and Hevea brasiliensis,R. communis, E.schimperi and E.esula)*, Malvales, Myrtales and Sapindales of, Solanales of asterids and the early diverging eudicot order Ranuncuales *(Thalictrum coreanum, Megaleranthis saniculifolia* and *Ranunculus macranthus).* (Raman & Park, 2015; Park et al., 2015; Tangphatsornruang et al., 2011). The gene *rpl32 is also missing in several angiosperms, including Thalictrum coreanum*, *Dianthus superbus* and *Populus* (Steane,

2005; Okumura, 2006; Park et al., 2015; Raman &Park, 2015). In the case of *Populus rpl32*, the nuclear genome has been documented.

1.3.3. Comparison to the Plastomes of other Euphorbiaceae species

The *Euphorbia schimperi* plastome size is similar to other Euphorbiaceae. Overall, *Jatropha curcas* has the largest genome (163,856 bp) while *Euphorbia schimperi* has the smallest (159,463 bp). The LSC of *Euphorbia schimperi* is the smallest (88,900 bp) while *Jatropha curcas* has the largest LSC (91,756 bp). The shortest SSC is in *Euphorbia esula (*17,023 bp) while the largest SSC occurs in *Ricinus communis (*18,796 bp). The overall GC content in all the species was similar and the number of rRNA genes (4) in each IR copy was the same for all species of Euphorbiaceae. There was a deviation in the total number of protein coding genes with *Euphorbia schimperi* and *Euphorbia esula* having a total of 85 protein coding genes while *Hevea brasiliensis* and *Jatropha curcas* had the fewest protein coding genes at 78. The *atpF* intron has been reported to be lost in *Hevea brasiliensis* and *Manihot esculenta* (Daniell at el., 2008). In case of *infA*, it is non-functional and nearly missing in *E. esula* and *E. schimperi, J. curca, M. esculenta*, *R. communis*. The pseudogene *rps16* was reported in *J. carcus* (Asif at el., 2010). The differences between the genomic characteristics of the aforementioned species are provided in **Table 1.1 and Figure 1.4.**

1.3.4. Boundaries of LSC, SSC and IR regions

In plastomes, the (IR) remains the most conserved region despite its expansion and contraction in various angiosperm lineages (Ruhlman and Jansen 2014). **Figure 1.2** presents a detailed comparison of the IR-SSC and IR-LSC boundaries among seven Euphorbiaceae plaastomes compared to *Nicotiana tabacum*. The *rps19* pseudogene located at the LSC/IRb boundary varies in size from 37 bp in *E. schimperi*, 92 bp in *M. esculenta,* and 82 bp in *H. brasiliensis.* The entire *rps19* coding region occurs in the LSC of *E. esula* and *J. curcas*, whereas the gene is duplicated in the IR in *V. fordii* and R. *communis*. The IRb/SSC border expands into *ycf1* in *E. schimperi* (10 bp), *V. fordii* (14bp), *R. communis* (19 bp) and *E. esula* (25 bp). The length of the *ycf1* pseudogene is approximately 1,222 bp in *V. fordii*, 1418 in *E. esula,* 2,200 bp in *J. curcas,* 1,177 bp in *R. communis* and 1,388 bp in *E. schimperi.* Furthermore, *ndhF* was located entirely in SSC in *E. schimperi*, *J. curcas*, *R. communis*, *E. esula, V.fordii* and *H. brasiliensis*. The location of *trnH*-GUG in the LSC varied in *E. schimperi*, *M. esculenta*, *J. curcas* and *H. brasiliensis* due to expansion/contraction of IRa. Overall, there is a slightly different of IR boundary regions of the *E. schimperi* plastome compared to other Euphorbiaceae.

Alignments of eight Euphorbiaceae plastomes indicate all have the same gene order and orientation of syntenic blocks (**Figure 1.3).** This is with the exception of *H. brasiliensis,* which has an inversion of 30,000 bp in the LSC between the *trnE-UUC and trnR-UCU* genes (Tangpatsornruang et al., 2011).

1.3.5. Phylogenetic analysis

The phylogenetic tree revealed that *E. schimperi* is sister to the *E. esula* with 98% bootstrap support (**Figure 1.4**). They are quite similar in term of gene loss as well with both species missing *rpl32* and *rps16.* This phylogenetic is similar to those of previous studies of Euphorbiaceae (Asif, at el., 2010) (Daniell et al., 2008) (Rivarola et al., 2011) (Li et al., 2017)) with *M. esculenta* and *H. brasiliensis* forming a clade that is sister to *Euphorbia* with 76% bootstrap support. The monophyly of *M. esculenta* and *H. brasiliensis is* also supported by an *atpF* intron loss. The sister relationship of V. *fordii* and *J. curca* is supported by 75% bootstrap and the loss of *rps16* (Li, 2017).

1.4. Conclusion

To summarize, the plastome size and arrangement are similar to other land plants, with the exception of the loss of *rps16*, *rpl32* and *infA* genes. The phylogenetic analyses by ML method show that *E. schimperi* forms a strong relationship with *E. esula* (98% BS). The availability of the medicinal plant *E. schimperi* plastome sequence can facilitate further investigations medicinally important species. *Euphorbia schimperi* plastome will be valuable in the assessment of evolutionary relationships among plant species and expand knowledge of the phylogenetic distribution of *atpF* intron loss and losses *rpl32*, *rps16*, and *infA* genes.

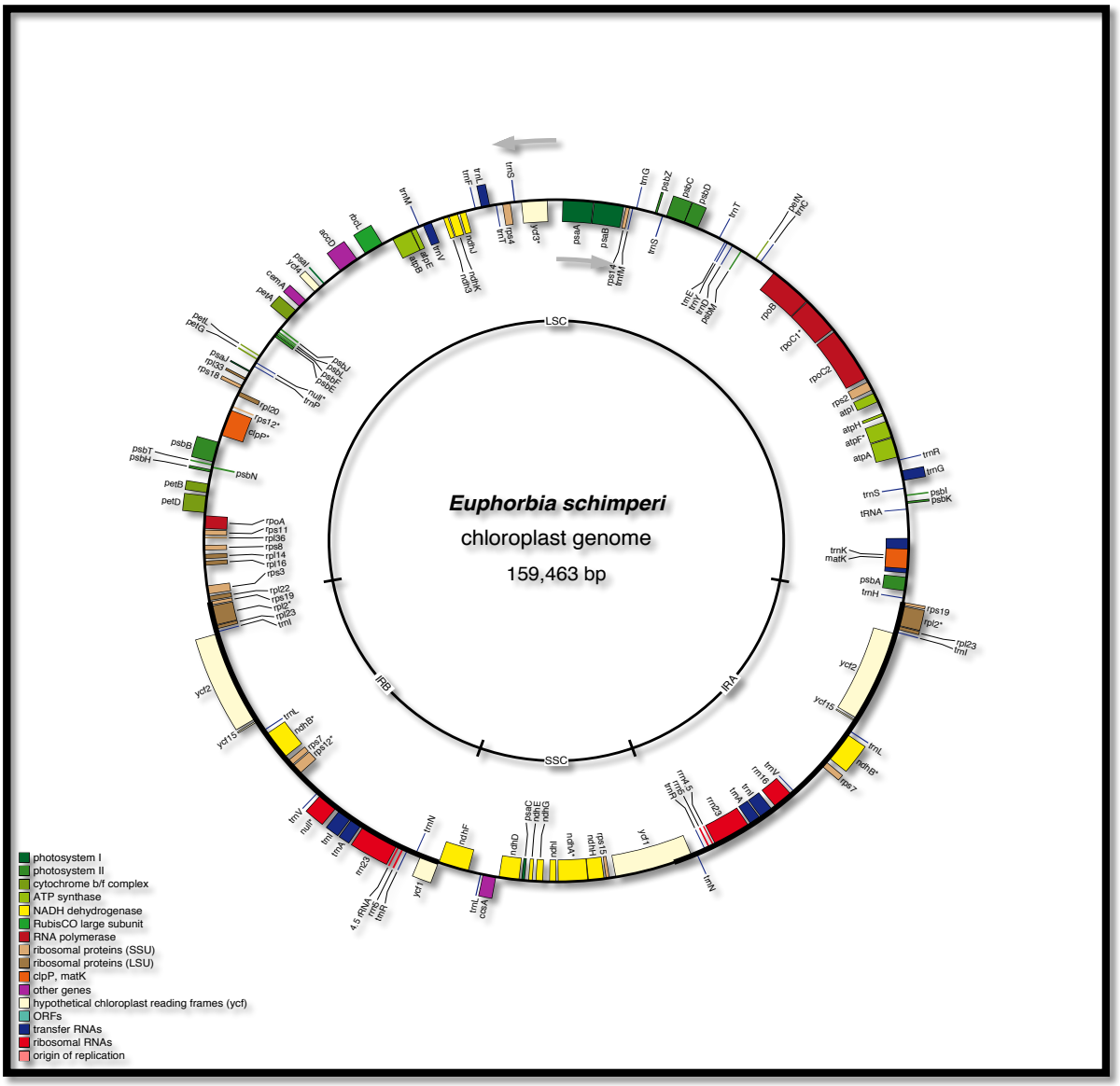

Figure1.1: Plastome map of *E. schimperi*. The gene drawn outside of the circle are transcribed clockwise, while those inside are counterclockwise. Small single copy (SSC), Large single copy (LSC), and inverted repeats (IRa, IRb) are indicated. The grey arrows show direction of transcription.

Genome featur	Euphorbia schimperi	Euphorbia esula	Hevea brasiliensis	Jatropha curcas	Manihot esculenta	Ricinus communis	Vernicia fordii
genome size	160512bp	160512bp	161191bp	163856bp	161453bp	163161bp	161528bp
LSC length	88900	90840 17023	89209bp	91756bp	89295bp	89651bp	89132bp
SSC length	17305bp	17023	18362bp	17852bp	18250bp	18796bp	18758bp
IR length	26629bp	26344	26810bp	27124bp	26954bp	27347bp	26819bp
GC content %	35.6	35.6	35.7	35.4	35.9	35.7	36
Number of genes	118	119	128	130	128	129	135
Genes duplicated in IR	19	17	16	17	16	17	21
Protein-coding genes	84	85	78	78	83	83	81
tRNA genes	30	30	30	28	30	38	29
rRNA genes	8	8	8	8	8	8	8

Table1.1: Comparison of general features of seven Euphorbiaceae plastomes.

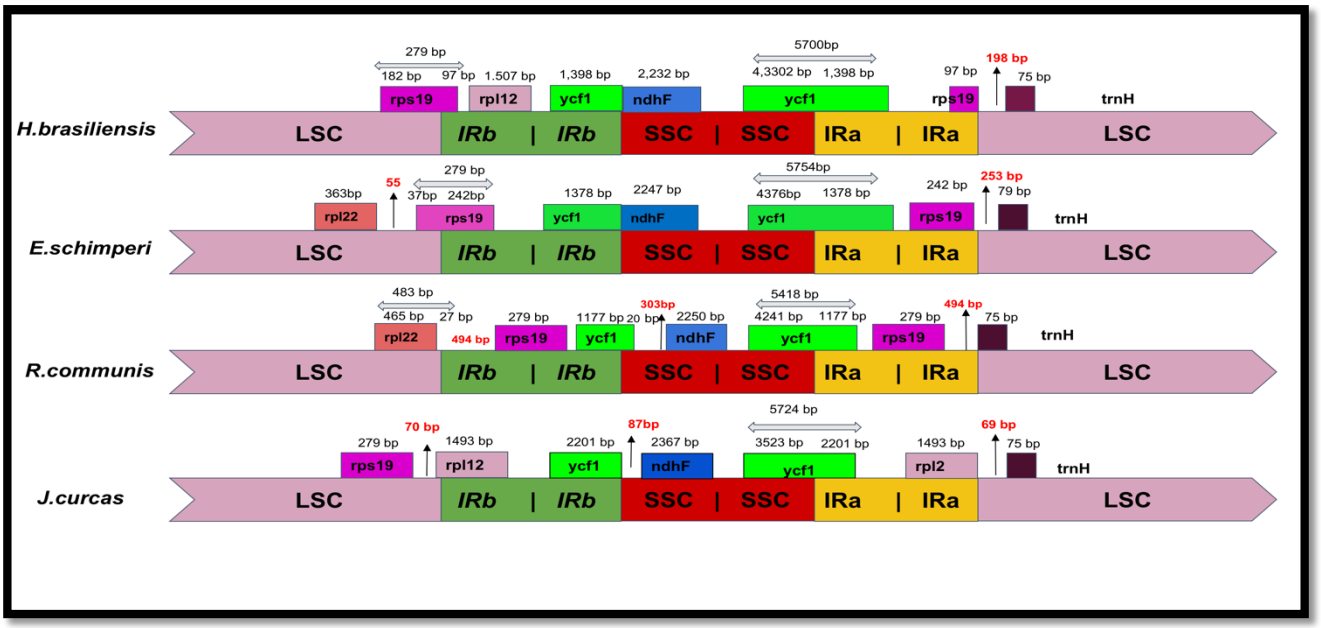

Figure 1.2: Comparison of chloroplast genome border of LSC, SSC, and IRs among seven species of Euphorbiaceae and Tobacco.

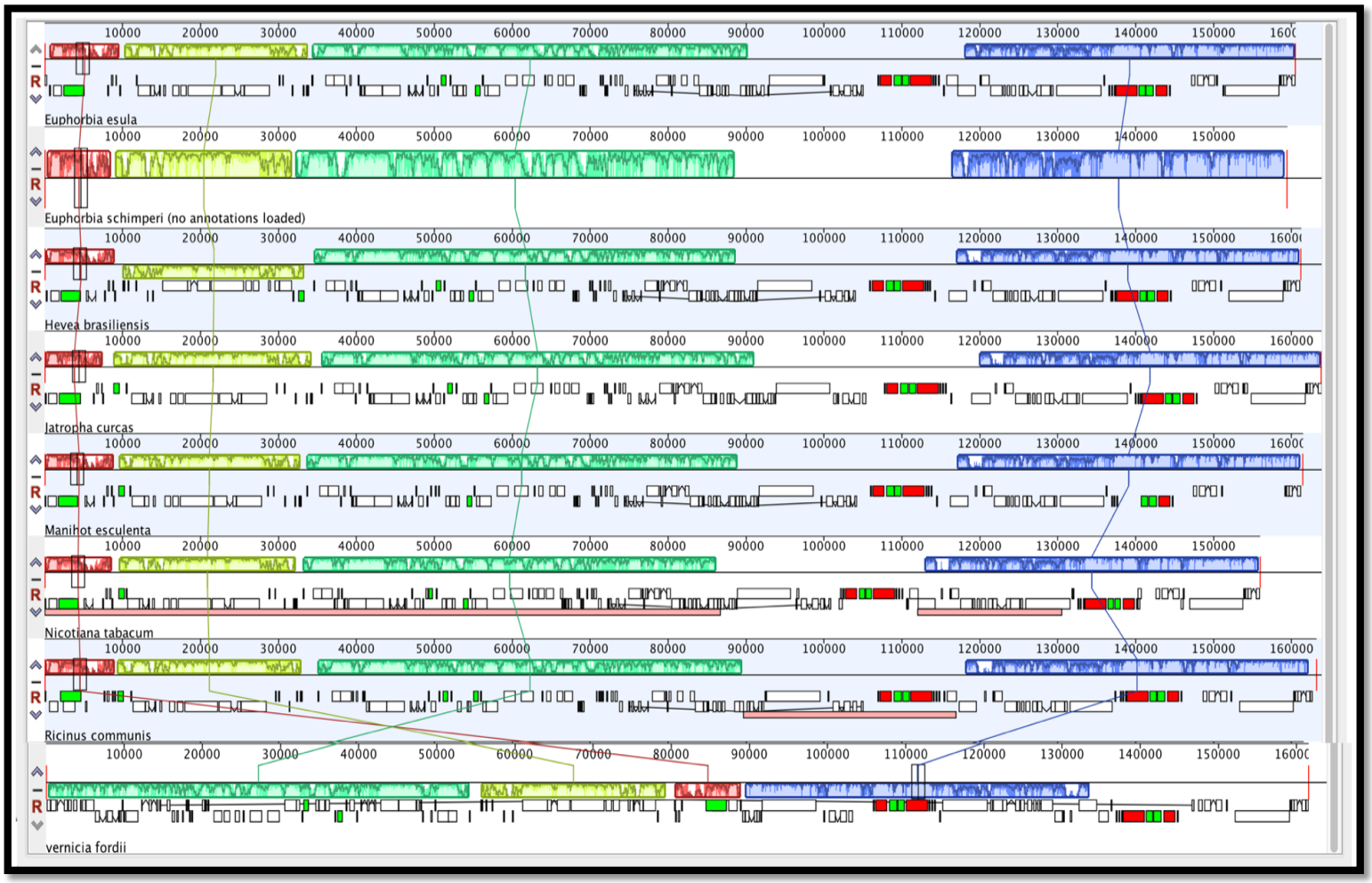

Figure 1.3: Synteny and rearrangements detected in Euphorbiaceae plastomes using the Mauve multiple- genome alignment program.

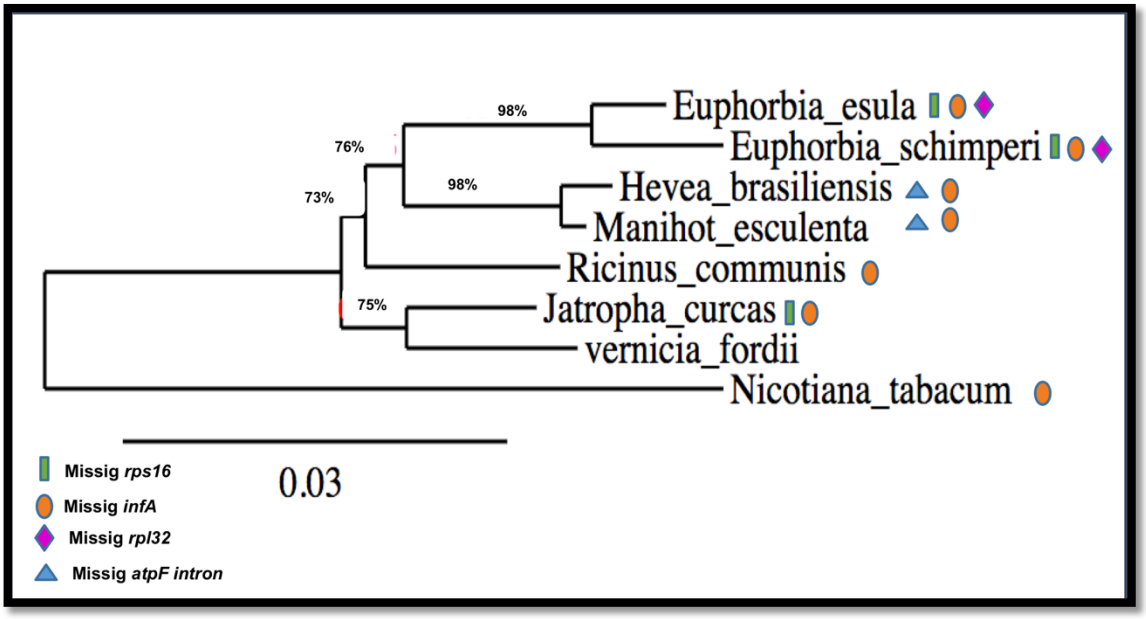

Figure 1.4: Phylogenetic tree of Euphorbiaceae inferred from the plastid genome sequences.

Chapter 2: Characterization of Genes involved in Producing Anti-Cancer Compounds in the *Euphorbia schimperi* Transcriptome

2.1. Introduction

Euphorbia schimperi is a medicinal plant in the family Euphorbiaceae. Many species in the family are sources of nourishment while others are helpful for their oils and waxes. Some are a source of medicinal drugs, but others are dangerous and poisonous (Augustyn et al., 2018). Species of Euphorbiaceae cannot grow in arctic and cold alpine regions but most of them grow in tropical and temperate regions. Also, the family has both annual and perennial herbs, woody trees, and a few climbers. Four Euphorbiaceae species that are important economically are described below. *Manihot esculenta* (cassava) is the third-largest source of food carbohydrates in the tropics, after rice and maize (Claude and Denis, 1990). *Ricinus communis* (caster bean) seed is the source of castor oil, which has a wide variety of uses (Rivarola et al., 2011). *Jatropha curcas* seed is processed to produce a high-quality biodiesel fuel (Achten, 2007). *Hevea brasiliensis (*rubber tree**)** is the most economically important member of the genus *Hevea*. The milky latex provides an important primary source of natural rubber (Tangphatsornruang et al., 2011). These plants are economically important and have been shown to have some medicinal activities due to the presence of isoprenoids (Ahmed et al., 2016).

Euphorbia schimperi is a member of a diverse genus of flowering plants commonly called spurge. The flowers are distinguished by cup-shaped clusters called

cyathia, which are composed of a single pistil and stamen. They have simple, alternate leaves along the stem, which has a milky latex sap to protect the plant from herbivores. The milky sap has been used to treat cancer (Jade, 2012; Pharm et.al, 2017). Some triterpenoids isolated from *Euphorbia* species have antimicrobial and anti-cancer properties, such as *E. thymifolia*, *E. hirta* and *E. cheiradenia* (Mothana et al., 2009). Moreover, *Euphorbia schimperi* is also known to produce seven compounds that possess anti-cancer properties. These compounds include triterpenoids like cycloart-23-en-3β, 25-diol, α-amyrin, and cycloart-25-en-3β 24-diol as well as steroid glycosides, flavonoids and phenylpropanoids (Abdel-Monem et al., 2008). Other compounds produced by *Euphorbia schimperi* also possess medicinal applications (**Table 2.1).**

Compound Class	Compound Name	Medicinal Effect
Triterpene	Cycloart-25-en-3β,24-diol	Has anti-proliferation properties that make it suitable for use in cancer treatment (Ayatollahi et al., 2011).
	Cycloart-23-en-3β,25-diol	Exhibits diverse anti-bacterial activity and strong activity against yeast fungi (Badole et al., 2011).
	α-Amyrin	When combined with β-Amyrin, the compound is applicable to the development of a drug that would facilitate the effective treatment of atherosclerosis and diabetes (Santos et al., 2012).
Flavonoids	Luteolin	Has an anti-inflammatory effect thus, is used in cancer treatment (Lin et al., 2008).
	5-kaempferol	Has many pharmacological uses including microbial, anti-inflammatory and anti-cancer uses among others (Calderon-Montano et al., 2011).
Phenylpropanoids	Scopoletin	Mainly used as an anti-fungal agent among other pharmacological activities (Gnonlonfin et al., 2012).
Phytosterols	7-β-sitosterol-β-D-O-glucoside	It is used to lower the level of cholesterol (Saeidnia et al., 2014).

Table 2.1: Medicinal applications of the seven compounds of *E. schimperi*.

Having information about a transcriptome can enhance the medicinal use of a plant because it allows analysis of plants to determine their qualities through mechanisms such as molecular biology and phytochemistry. This study will discuss *E. schimperi* in detail with the goal of determining the genes that are involved in biosynthetic pathways responsible for the production of *E. schimperi* anticancer secondary compounds.

2.2. Methods

2.2.1. Plant material and RNA isolation

The fresh leaf tissue of *Euphorbia schimperi* was collected from the Arid Land greenhouse in Tucson, Arizona. A voucher specimen of *E. schimperi* was deposited in TEX herbarium. RNA was extracted from 0.25 g of fresh leaves using the RNeasy Plant Mini Kit following the manufacturer's instructions (Qiagen, CA, USA). Using DNase digestion, RNA was treated to eliminate any remaining DNA based on the enzyme protocol (Fermentas #EN0521, 1 unit/ul). The 50ul RNA sample was combined with 35ul water, 10ul DNase, 10x buffer, and 5ul DNase for a total volume of 100ul. After incubation for 1h at 37°C, the DNase was removed through the microcolumns and then cleaned with RNA Clean & Concentrator-25 (Zymo Research). The RNA sample, containing 260/280 ratios from 1.9 to 2.1, 260/ 230 ratio from 2.0 to 2.5, and RNA integrity number (RIN) > 8.0 was used for cDNA library preparation analysis.

2.2.2. Transcriptome *de novo* assembly

Standard RNA-Seq library preparation and sequencing via Illumina HiSeq 4000 were carried out in the Genome Sequencing Analysis Facility (GSAF, the University of Texas at Austin). The minimum expected number of reads was 30 million paired-end sequences (2 x 150) while the targeted reads were 36 million. The quality of raw FastQC reads was observed using the FastQC tool to get high quality *de novo* transcriptome sequence data. Once the quality of the RNA reads was good the reads were no longer subjected to any filtering methods. A *de novo* assembly of reads into the transcripts was performed using Trinity (Grabherr, *et al.,* 2011*)*. Trinity sequentially integrates

Inchworm, Chrysalis, and Butterfly modules to process a large number of RNA-Seq reads. This has been used to partition the sequence data into different individual de Bruijn graphs, which represent the transcriptional complexity at a given gene or locus (Hass, 2013; Liu et al., 2013).

2.2.3. Quality Assessment and Annotation

To validate the *de novo* assembly, read remapping was conducted by using two software packages, Bowtie 2 (Langmead, 2012) and BUSCO (Rebert et al., 2017; Felipe et al., 2015). Bowtie 2 index was created for the data and counted the number of reads to map to the transcriptome. BUSCO v3.0.2 was carried out using the Embryophyta and Eukaryota database. The BUSCO assessment provided the quantitative measures to identify the completeness of the transcriptome based on evolutionarily informed expectations of the gene content from near-universal single-copy orthologs selected from OrthoDB v9 (Rebert et al., 2017; Felipe et al., 2015). The final transcriptome assembly for *E. schimperi* was annotated using the BLASTX searches against the protein database (SwissProt) and the predicted *Arabidopsis thaliana* proteome (Tair V10, http://arabidopsis.org) using the BLAST settings BLASTX, report 1 hit, e-value $1E^{-5}$.

2.2.4. Gene Search

Biosynthetic pathways were examined for the six compounds (triterpenoids; cycloartenol, and α-amyrin, in addition to steroid glycoside (β-sitosterol-β-D-O-glucoside), phenylpropanoids (scopoletin), and flavonoids (luteolin and kampferol) using MetatCyc (Caspi et al., 2014), a highly curated metabolic pathway database. The initial steps for gene searching included checking the final stage of the compound production in

the biosynthetic pathway by identifying the gene that creates the enzyme to produce the compound. Second, there were six required genes to produce six compounds in the four species and six biosynthetic pathways (**Table 2.1**). Third, the transcriptome nucleotide sequences were converted to amino acids using TransDecoder as cited in https://github.com/TransDecoder/TransDecoder/wiki. Fourth, 10 closely related species with transcriptome sequences from phytozome (*Arabidopsis thaliana, Manihot esculenta, Populus trichocarpa, Ricinus communis, Salix purpurea, Linum usitatissimum, and Glycine max*), the Medicinal Plant Genomic Resource (*Hypericum perforatum*) and from NCBI (*Hevea brasiliensis* and *Jatropha curcas*) were downloaded. Fifth, a database of the 10 transcriptomes closely related to *Euphorbia schimperi* was created using makeblastdb command. Sixth, all the protein sequences for all the six genes were download from Uniprot and placed them in one file. Then for the last two steps, the six query representative genes were blasted to the 10 species database to identify the matches for every representative gene using blast command setting (blastp and e-value 1e-2). Without specifying the maximum target sequence, the goal was to identify all the possible matched genes. The result file was filtered, which involved deleting sequences for all 10 species that had ≥ 50% percent identity. Then, the six query representative genes were blasted to the *E. schimperi* translated transcriptome and the result file was filtered at ≥50% percent identity to identify matching sequences from *E. schimperi* transcriptome.

2.2.5. Gene Family Analysis

The overall strategy for gene family analyses is summarized in Figure 2.1. Sequences of the six representative genes that have a match with the 10 species database and *E. schimperi* transcriptome at ≥ 50% percent identity were extracted. Then, protein sequences for the genes that were used as references and involved in biosynthetic pathways of the compounds (LUP2 gene, a-Amyrin compound*,* in *Arabidopsis thaliana)*, (AT4G34050 (CCOAOMT1 gene, Scopoletin compound in *Arabidopsis thaliana)*, and (AY230247 (FNSI) gene, Luteolin compounds in *Petroselinum crispum)*, *UGT80A2gene,* b-SITOSTEROL-b-D-O-glicoside compounds in *Arabidopsis thaliana*), RcCAS gene, cycloartenol compound in *Rhizophora stylosa* , FLS1 gene, kaempferol compound in *Fagopyrum tataricum*) were aligned to the sequence that have ≥50% identity matches from 10 protein species database and *E. schimperi* transcriptome using MUSCLE (Edgard, 2004). Sequences covering < 50% of the length each gene were removed and sequences extended beyond the length of the reference gene were trimmed. After the removal and filtering of the sequences, alignments were carried out again with MUSCLE for maximum likelihood (ML) analysis with RAxML v8 (Stamatakis, 2014) in Geneious. MUSCLE settings included eight iterations with sequences classified based on similar features using the kmer 6_6 for the first iteration and pctid_kimura distance for all subsequent iterations and the unweighted pair group method with arithmetic mean (UPGMB) clustering method. The ML analyses used the GAMMA GTR protein model and rapid bootstrap algorithm while 1000 replicate trees were the number of starting trees or bootstrap replicates (Figure 2.1). Trees were rooted using *Arabidopsis*

thaliana in the Brassicales in the rosid II clade, which is sister to the rosid I clade that included all other species examined.

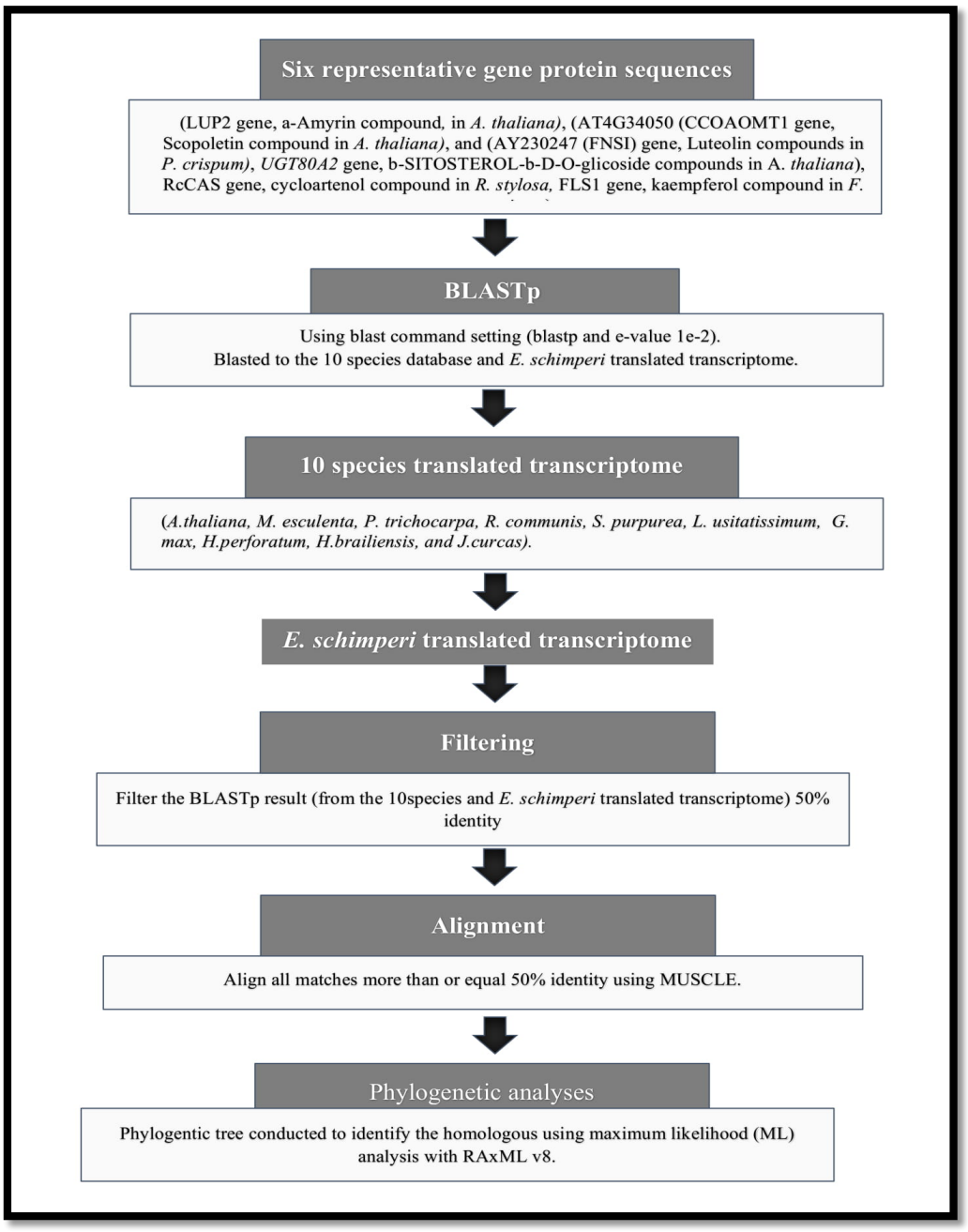

Figure 2.1: Summary of steps involved in generating gene family trees.

2.3. Results

2.3.1. Assembly of *Euphorbia schimperi* transcriptome

The sequenced Illumina libraries yielded 80,916,952 million reads. Assembly used 25-mer in Trinity, which is recommended by its authors (Hass. J et al., 2013). The total reads used and total assembled contigs and N50 statistics are indicated in **Table 2.3**. The mapped read coverage to the assembly using Bowtie2 (Langmead, B ,2012) is 90.26% and half of the transcriptome is in contigs larger than or equal the 1,127bp (N50) contig size. Busco indicated that the transcriptome covered 87% of 100 species in Eukaryota and 72.30% of 30 species in embryophyta (**Table 2.4**). The statistics of trinity assembly data are in **Table 2.3**.

2.3.2. Annotation of the transcriptome assemblies

The final transcriptome assembly for the *E.schimperi* was annotated using a combination of evidence from sequence similarity searches to the SwissProt database and the predicted *Arabidopsis thaliana* proteome (TAIR10, http://arabidopsis.org). The statistics of translated trinity assembly data are identified in **Table 2.3**.

2.3.3. Genes involved in biosynthetic pathways

All six query genes with ≥ 50% percent sequence identity were found in *E. schimperi* by using blastp (Table 2.2). Phylogenetic trees were generated for six genes. *Camt4*(CCOAOMT1), the reference gene from *A. thaliana(ART)* was clustered with identical copy of A. *thaliana(ART)* and was homologous with 4 copies *of E. schimperi* with bootstrap support of 54% (Figure 2.2). *FLS1* genes from *Fagopyrum tataricumis* (Figure 2.3) and *FNS I* from *Petroselinum crispum* (Figure 2.4) grouped with *E.*

schimperi copies even though they are further away from each other. *LUP2* from *Arabidopsis thaliana,* has more orthologue copies (8 copies) in *E. schimperi* and clustered with other copies of *A. thaliana* (Figures 2.5). *RcCAS* from *Rhizophora stylosa and UGT80A2* from *Arabidopsis* have more orthologue copies (35 copies total) in *E. schimperi (Figure 2.6-2.7).*

2.4. Discussion:

The six gene family trees that were constructed from aligning the ≥50 % matches from 10 species and *E. schimperi* show expansion and diversification among homologs encoding the six biosynthetic pathways. LUP2 and *UGT80A2* from *A. thaliana and RcCAS* from *R. stylose* show more expansion than the other genes. These protein families have potential for testing the discovery of novel functions in Malpighiales. A similar study was conducted in *Rhazya stricta*, and showed a lineage-specific expansion and diversification among homologs encoding MIA pathway genes in Gentianales (Sabir et al., 2016). This study suggested that the expansion of *LUP2*, *UGT80A2* and *RcCAS* family genes might associated with alpha-Amyrin, 3β-hydroxy sterol glucosyltransferase, and cycloartenol synthase biosynthesis in *E. schimperi*. In a previous study, *H. brasiliensis* genome showed expansion of gene family associated with rubber biosynthesis (Lau at el., 2016). Expansion might be associates with the evolution of novel functions such as production of floral structures, induction of disease resistance, and adaptation to stress (Panchy et al., 2016). Additionally, important agronomic traits, such as grain quality, fruit shape, and flowering time have been reported recently in wheat (*Triticum aestivum*), cotton (*Gossypium hirsutum*), and soybean (*Glycine max*) due

to the whole-genome duplications (Panchy et al., 201). The expansion of gene family might assist a novel function in *E. schimperi*. Comparing *E. schimperi's* genes to the plants that have the same genes involved in the biosynthetic pathway, specifically *A. thaliana, R. stylose, F. tataricumis,* and *P. crispum* as well as to other related species in the same family whose transcriptomes are already known such as (*Ricinus communis* (RIC), (Manihot esculenta (MAE) and (*Jatropha curcas* (JAC) assist to identify the putative candidate genes in *E. schimperi* (*LUP2*, *CCOAOMT1*, (FNSI), *UGT80A2, RcCAS*, and *FLS1).*

Symbol	Full Name	Accession #	compounds	enzyme	Sources
LUP2	Amyrin synthase LUP2	Q8RWT0	a-Amyrin	**α-amyrin synthase / β-amyrin synthase**	*Arabidopsis thaliana*
(CCOAOMT1)	Caffeoyl-CoA O-methyltransferase 1	O49499	Scopoletin	**caffeoyl-CoA *O*-methyltransferase**	*Arabidopsis thaliana*
AY230247 (FNSI)	Flavone synthase I	Q7XZQ8	Luteolin:	**flavone synthase I**	*Petroselinum crispum*
UGT80A2	Sterol 3-beta-glucosyltransferase	Q9M8Z7	b-SITOSTEROL-b-D-O-glicoside	**3β-hydroxy sterol glucosyltransferase**	*Arabidopsis thaliana*
RcCAS	Terpene cyclase/mutase family member	A7BJ35	Cycloartenol	**cycloartenol synthase**	*Rhizophora stylosa*
FLS1	Flavonol synthase	F5BR59	Kaempferol	**flavonol synthase**	*Fagopyrum tataricum*

Table 2.2: Query sequences used in homolog search.

Total length of sequence	37478499bp
Total number of sequence	133847
N50	349
Max contig length	14119bp
Mean contig length	280
Total GC count	4649bp
GC%	8%

Table 2.3: Statistics of trinity assembly data

Total length of sequence	37478499bp
Total number of sequence	133847
N50	349
Max contig length	14119bp
Mean contig length	280
Total GC count	4649bp
GC%	8%

Table 2.4: Statistics of translated assembly data

Data	Number of species	Complete Buscos of *E.sch*	missing busco of *E.sch*
Eukaryota	100	87.10%	1.30%
Embryophyta	30	72.30%	14.90%

2.5: Quality assessment of using BUSCO

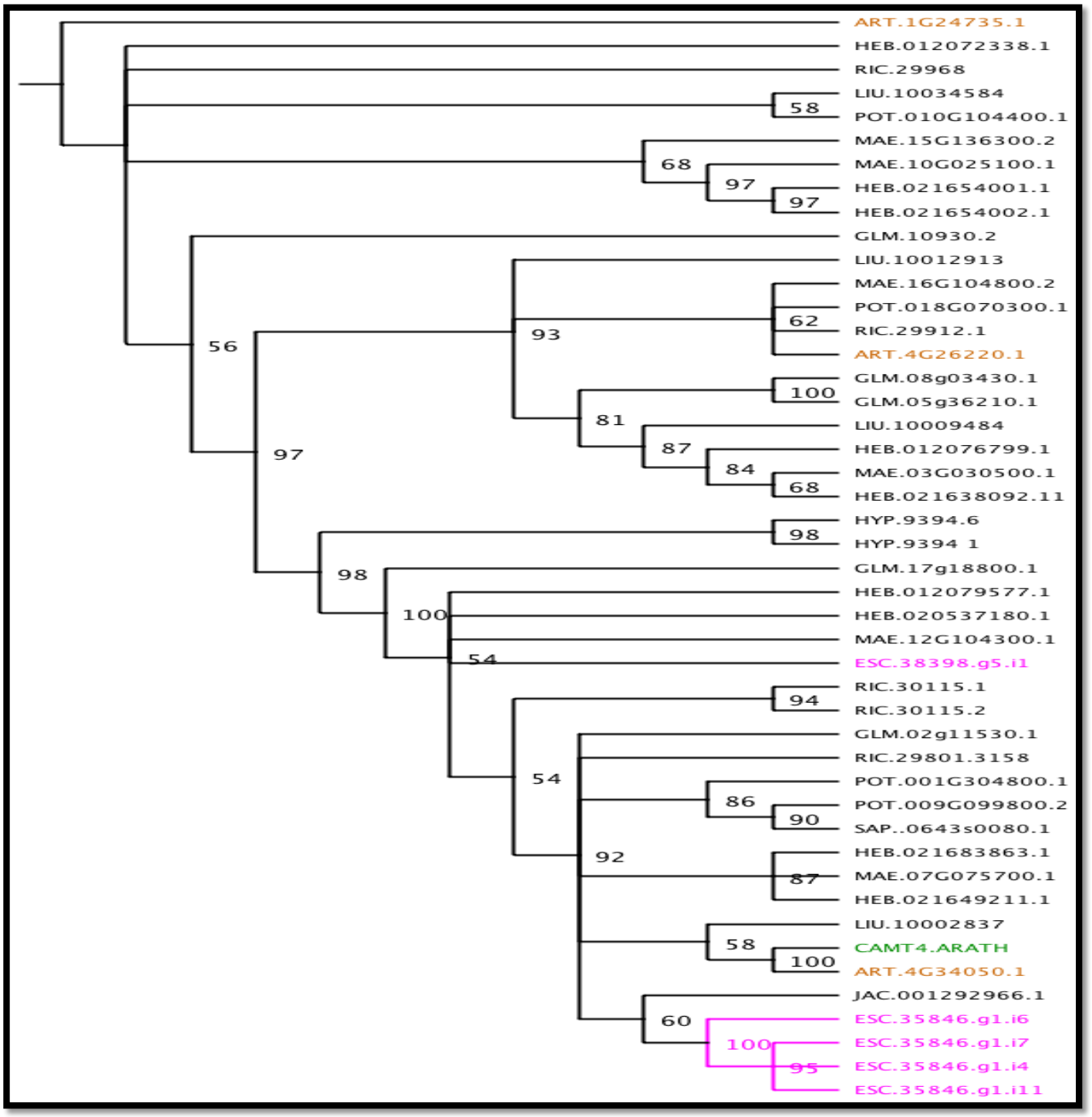

Figure 2.2: Phylogenetic tree of (*CCOAOMT1)* gene in (*Arabidopsis thaliana(Art*) green color, and 10 closely related species (*Ricinus communis* - RIC), (*Manihot esculenta* (MAE), (*Populus trichocarpa* (POT), (*Salix purpurea* (SAP), (*Linum usitatissimum* (LIU), (*Glycine max (*GLM), (*Hypericum perforatum* (HYP), (*Hevea brasiliensis* (HEB), (*Jatropha curcas* (JAC) and including (*Arabidopsis thaliana(Art*) transcriptome (brown color) and *Euphoriba schimperi* (ESC) copies (pink color).

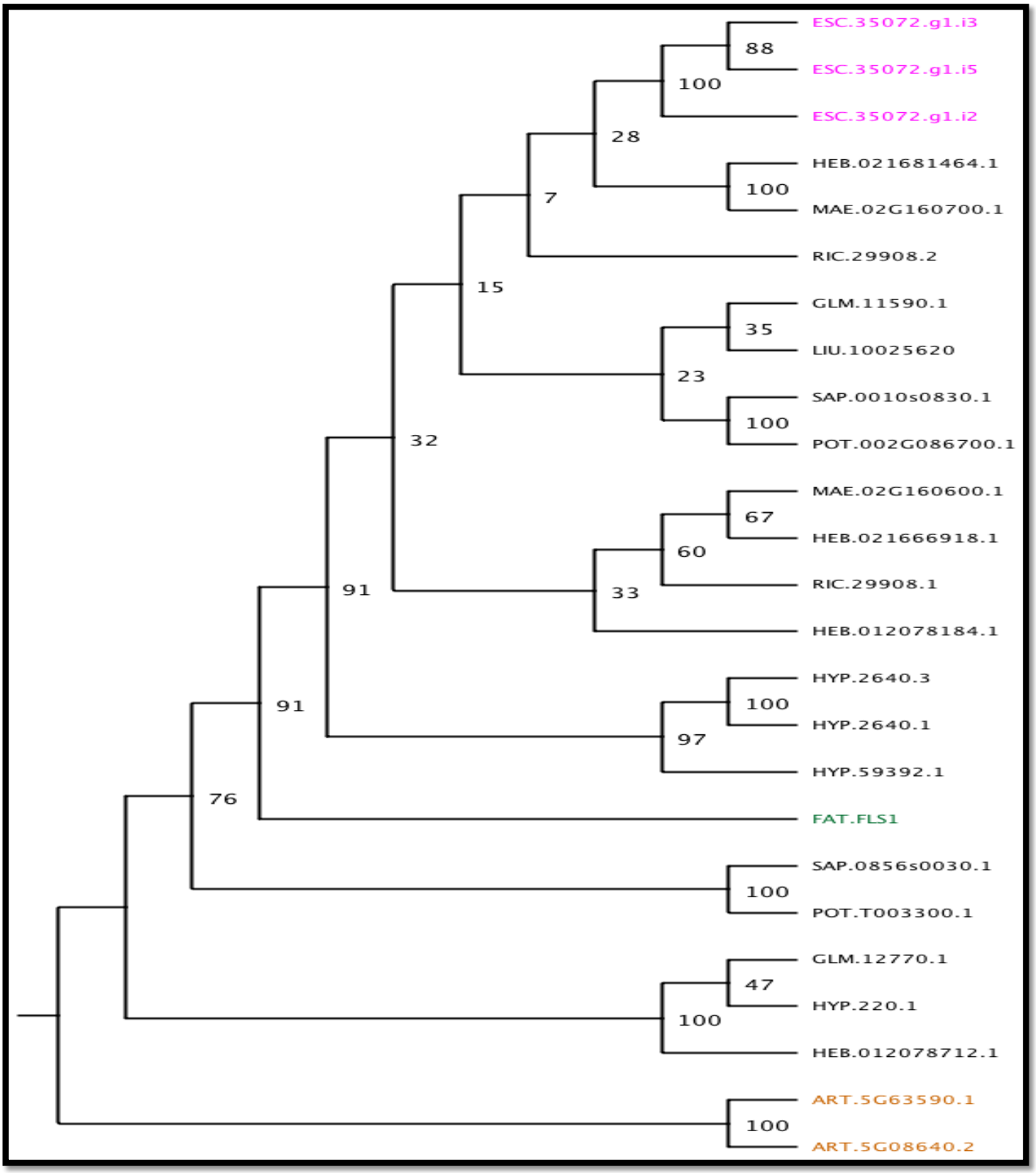

Figure 2.3: Phylogenetic tree of (***FLS1***) gene in (*Fagopyrum tataricurm*) (green color), and 10 closely related species (*Ricinus communis* - RIC), (*Manihot esculenta* - MAE), (*Populus trichocarpa* - POT), (*Salix purpurea* - SAP), (*Linum usitatissimum* - LIU), (*Glycine max* - GLM), (*Hypericum perforatum* - HYP), (*Hevea brasiliensis* - HEB), (*Jatropha curcas* - JAC), (*Arabidopsis thaliana - Art*) transcriptome orange color) and *Euphorbia schimperi* (ESC) copies (pink color).

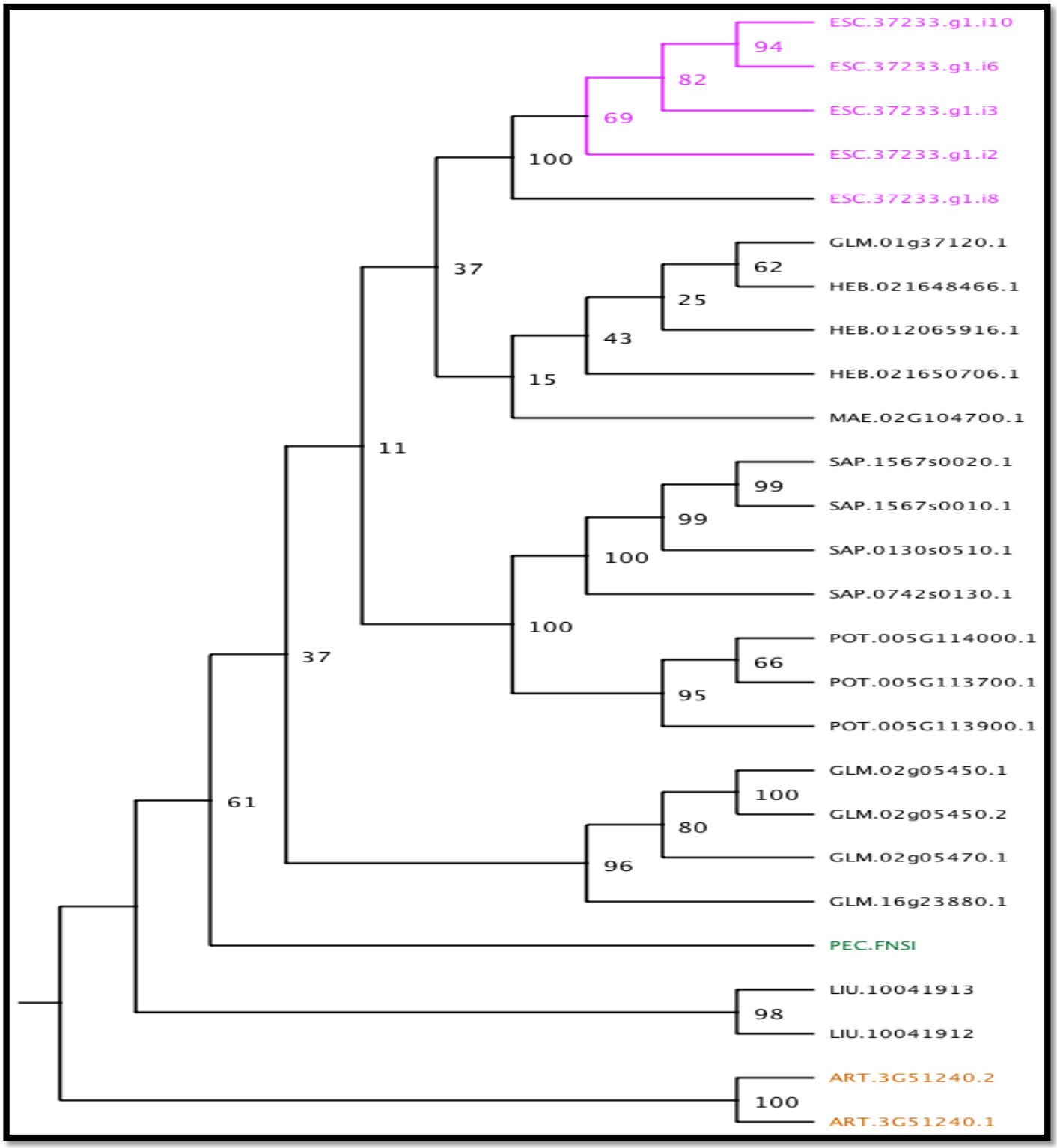

Figure 2.4: Phylogentic tree of (***FNSI***) gene in (*Petroselinum crispum*) (green color),10 closely related species (*Ricinus communis* (RIC), (Manihot esculenta (MAE), (Populus trichocarpa (POT), (Salix purpurea (SAP), (*Linum usitatissimum (*LIU), (*Glycine max (*GLM), (*Hypericum perforatum* (HYP), (*Hevea brasiliensis* (HEB), (*Jatropha curcas* (JAC), (*Arabidopsis thaliana(Art*) transcriptome (orange color) and *Euphoriba schimperi* (ESC) copies (pink color).

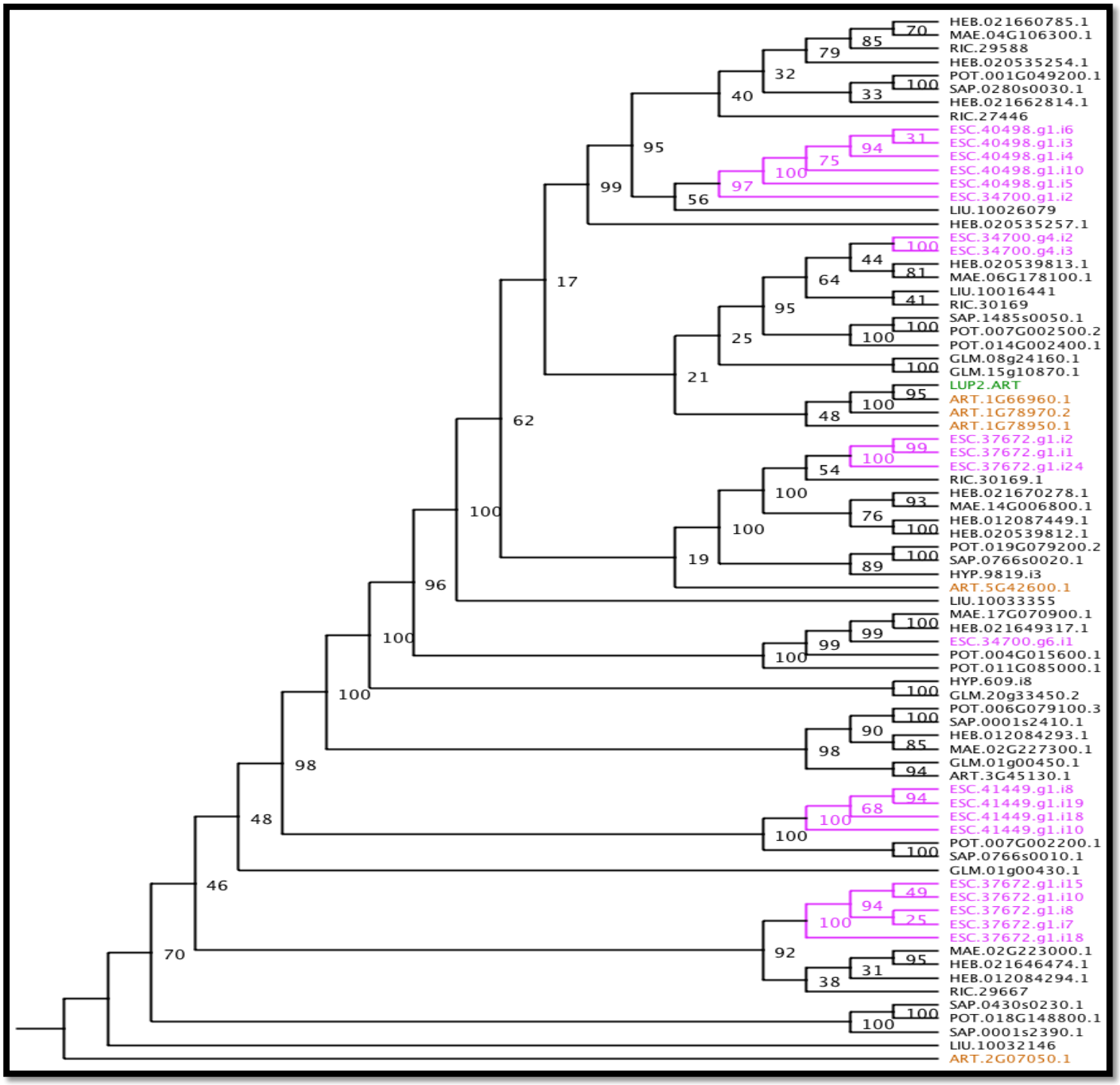

Figure 2.5: Phylogentic tree of (**LUP2**) gene in (*Arabidopsis thaliana(Art*) (green color), 10 closely related species (*Ricinus communis* (RIC), (Manihot esculenta (MAE), (Populus trichocarpa (POT), (Salix purpurea (SAP), (*Linum usitatissimum (*LIU), (*Glycine max (*GLM), (*Hypericum perforatum* (HYP), (*Hevea brasiliensis* (HEB), (*Jatropha curcas* (JAC) ,(*Arabidopsis thaliana(Art*) (brown color) transcriptome and *Euphoriba schimperi* (ESC) copies (pink color).

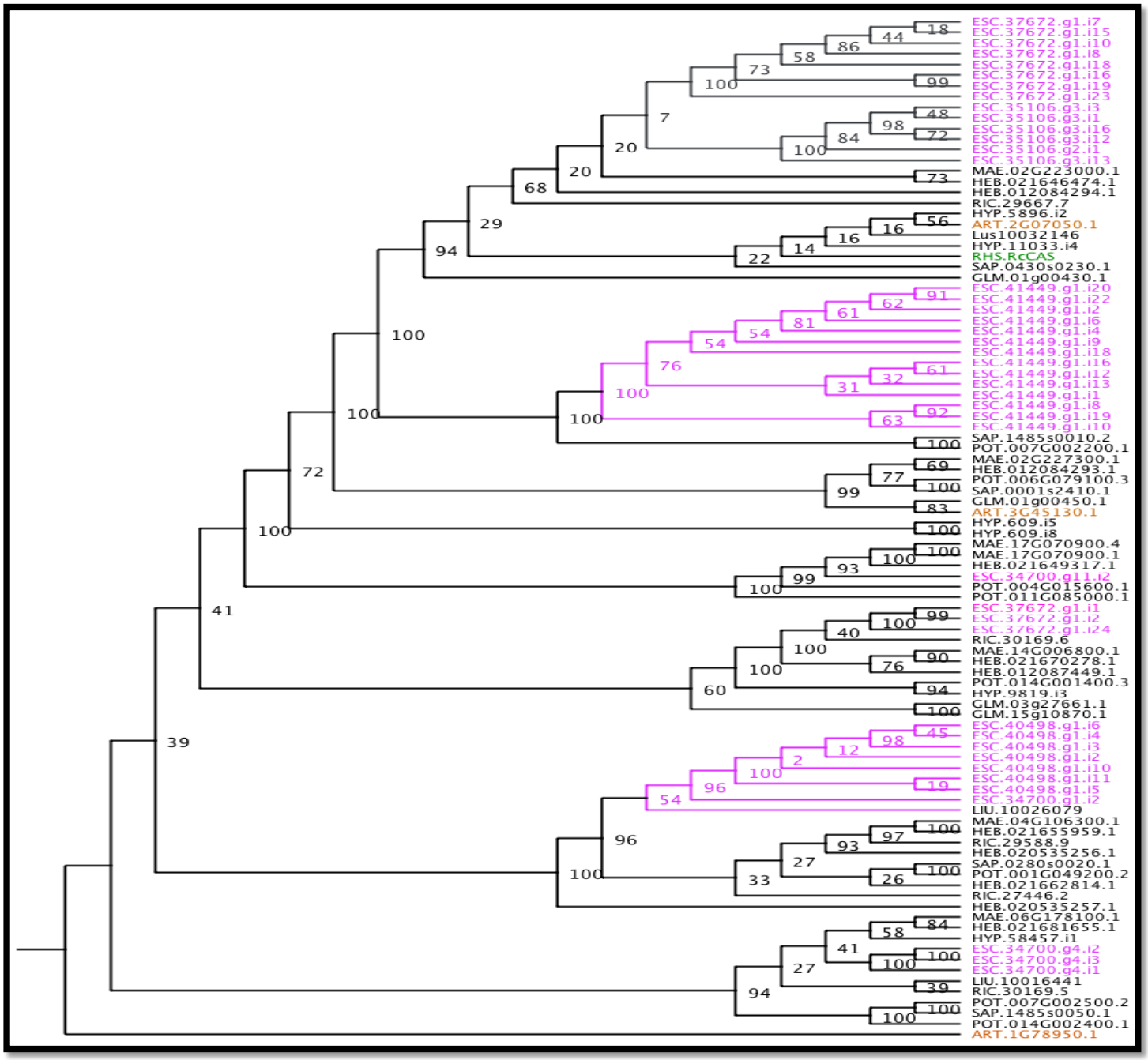

Figure 2.6: Phylogentic tree of (**RcCAS**) gene in (*Rhizophora stylosa)* (green color), 10 closely related species (*Ricinus communis* (RIC), (Manihot esculenta (MAE), (Populus trichocarpa (POT), (Salix purpurea (SAP), (*Linum usitatissimum (*LIU), (*Glycine max (*GLM), (*Hypericum perforatum* (HYP), (*Hevea brasiliensis* (HEB), (*Jatropha curcas* (JAC), (*Arabidopsis thaliana(Art*) (brown color) transcriptome and *Euphoriba schimperi* (ESC) copies (pink color).

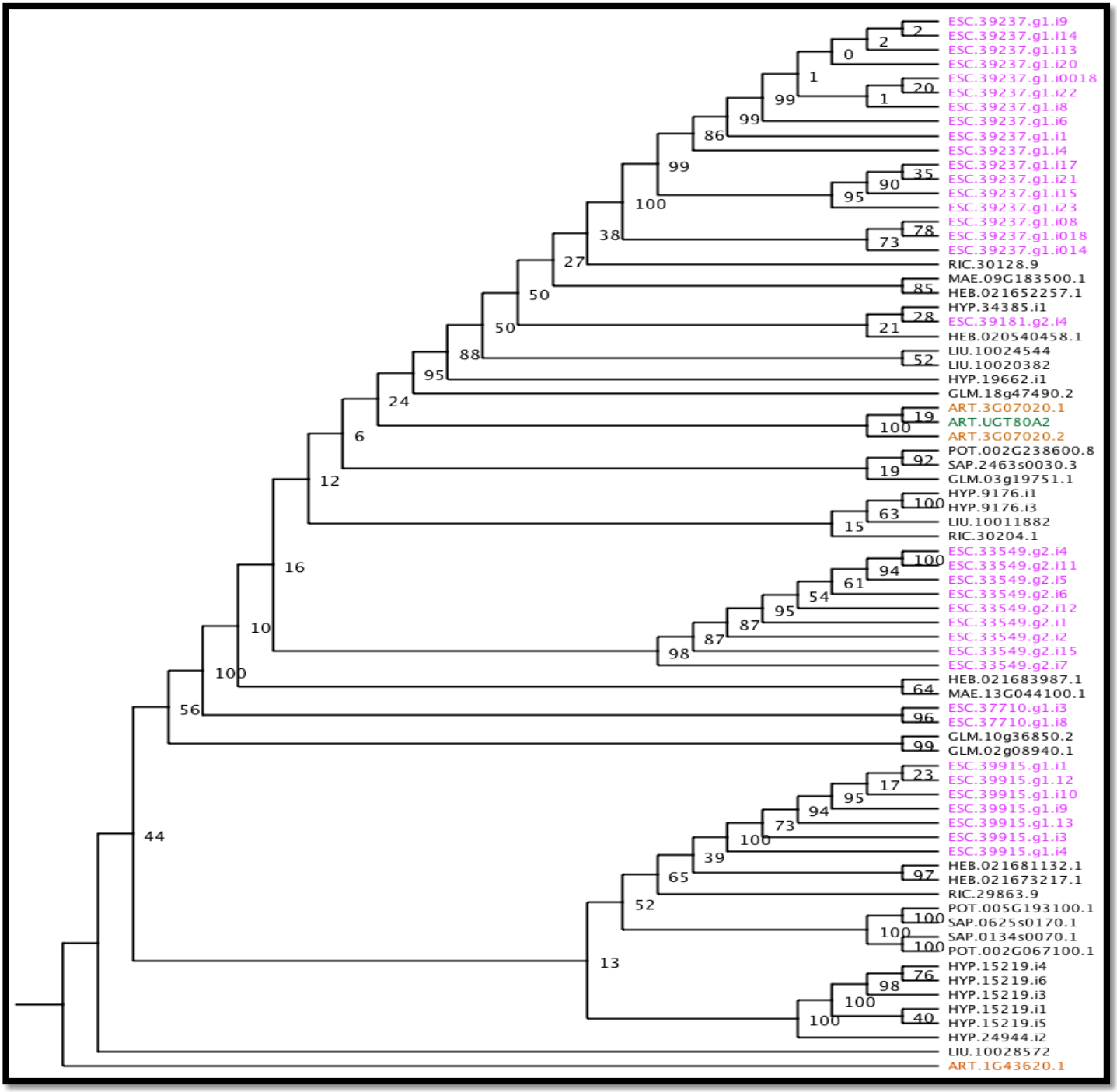

Figure 2.7: Phylogentic tree of (UGT80A2) gene in (*Arabidopsis thaliana(Art*) (green color), 10 closely related species (*Ricinus communis* (RIC), (Manihot esculenta (MAE), (Populus trichocarpa (POT), (Salix purpurea (SAP), (*Linum usitatissimum* (LIU),(*Glycine max* (GLM),(*Hypericum perforatum* (HYP),(*Hevea brasiliensis* (HEB),(*Jatropha curcas* (JAC) ,(*Arabidopsis thaliana(Art*) (brown color) transcriptome and *Euphoriba schimperi* (ESC) copies (pink color).

Bibliography

Abdel-Monem, A.R., Abdel-Sattar, E., Harraz, F.M. & Petereit, F. (2008). Chemical investigation of *Euphorbia schimperi* C. Presl. *Records of Natural Products*, *2*(2), 39-45.

Ahmed, S., Yousaf, M., Mothana, R. A., & Al-Rehaily, A. J. (2016). Studies on wound healing activity of some *Euphorbia* species on experimental rats. *African Journal of Traditional, Complementary and Alternative Medicines*, 13(5), 145–152.

Ayatollahi, A. M., Ghanadian, M., Afsharypuor, S., Mesaik, M. A., Abdalla, O. M., Shahlaei, M., & Mostafavi, H. (2011). Cycloartanes from Euphorbia aellenii Rech. f. and their antiproliferative activity. *Iranian Journal of Pharmaceutical Research: IJPR*, *10*(1), 105.

B. J. Haas, A. Papanicolaou, M. Yassour et al., "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis," Nature Protocols, vol. 8, no. 8, pp. 1494–1512, 2013.

B. J. Haas, A. Papanicolaou, M. Yassour et al., (2013) "*De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis," *Nature Protocols*, vol. 8, no. 8, pp. 1494–1512., doi:10.1038/nprot.2013.084.

B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," Nature Methods, vol. 9, no. 4, pp. 357–359, 2012.

Badole, S. L., Zanwar, A. A., Khopade, A. N., & Bodhankar, S. L. (2011). In vitro antioxidant and antimicrobial activity cycloart–23–ene–3β, -25–diol (B2) isolated from Pongamia pinnata (L. Pierre). *Asian Pacific Journal of Tropical Medicine*, *4*(11), 910-916.

Bhanot, A, R.S. Malleshappa, N. Noolvi (2011). Natural sources as potential anti-cancer agents: a review. *Bioinformatics*, published online June 9, 2015 10.1093/bioinformatics/btv351

Biomatters: Geneious R10 10.0.6 [http://www.geneious.com/].

Boyd, Jade 2012. A bit touchy: Plants' insect defenses activated by touch. Rice University.

Bramwell, D., Bramwell, Z. (2001). *Wild Flowers of the Canary Islands (2nd ed.). Madrid: Rueda.* ISBN 8472071294.

Buddensiek, Volker (2005): *Succulent* Euphorbia *plus* (CD-ROM). Volker Buddensiek Verlag.

Calderon-Montano, J., Burgos-Morón, E., Pérez-Guerrero, C., & López-Lázaro, M. (2011). A review on the dietary flavonoid kaempferol. *Mini Reviews in Medicinal Chemistry*, *11*(4), 298-344.

Caspi R, AltmanT, Billington R, Dreher K, Foerster H,Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, Krummenacker M, LatendresseM, Mueller LA, Ong Q, Paley S, SubhravetiP, Weaver DS, Weerasinghe D, Zhang P,Karp PD. 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucl Acids Res 42: D459-D471.

Cauz-Santos, L. A., Munhoz, C. F., Rodde, N., Cauet, S., Santos, A. A., Penha, H. A.,

Vieira, M. L. C. (2017). The Chloroplast Genome of Passiflora edulis (Passifloraceae) Assembled from Long Sequence Reads: Structural Organization and Phylogenomic Studies in Malpighiales. *Frontiers in Plant Science*, *8*, 334.

Chen C, Huang H, Wu CH (2017). Protein Bioinformatics Databases and Resources Methods Mol. Biol. 1558:3-39 (2017) .

Daniell, H., Wurdack, K. J., Kanagaraj, A., Lee, S. B., Saski, C., & Jansen, R. K. (2008). The complete nucleotide sequence of the cassava (Manihot esculenta) chloroplast genome and the evolution of atpF in Malpighiales: RNA editing and multiple losses of a group II intron. *Theoretical and Applied Genetics*, *116*(5), 723.

Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004; 14:1394–403.

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, (77), 772–782.

Edgar, R. C. *MUSCLE (2004): multiple sequence alignment with high accuracy and high throughput*. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

Fauquet Claude; Fargette Denis (1990*).* "African Cassava Mosaic Virus: Etiology, Epidemiology, and Control" (PDF). *Plant Disease.* **74**(6): 404–11. doi:10.1094/pd-74-0404.

Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, Evgeny M. Zdobnov; BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, Volume 31, Issue 19, 1 October 2015, Pages 3210–3212, https://doi.org/10.1093/bioinformatics/btv351.

Gnonlonfin, G. B., Sanni, A., & Brimer, L. (2012). Review scopoletin–a coumarin phytoalexin with medicinal properties. *Critical Reviews in Plant Sciences*, *31*(1), 4756

Grabherr, M. *et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome*. *Nature Biotechnol.* **29**, 644–652 (2011).

Guo, X, Liu J, Hao G, Zhang L, Mao K, Wang X, et al Plastome phylogeny and early diversification of Brassicaceae. BMC Genomics 2017; 18:176 pmid:28209119.

Hao, D. C., & Xiao, P. G. (2015). Genomics and evolution in traditional medicinal plants: Road to a healthier life. *Evolutionary Bioinformatics*, *11*, EBO-S31326.

Hong, S. Y., Cheon, K. S., Yoo, K. O., Lee, H. O., Cho, K. S., Suh, J. T., ... & Kim, Y. H. (2017). Complete chloroplast genome sequences and comparative analysis of Chenopodium quinoa and C. album. *Frontiers in Plant Science*, *8*, 1696.

Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA. 2007; 104:19369–19374. pmid:18048330.

Jansen, R.K. and T.A. Ruhlman. 2012. Plastid genomes of seed plants. In: Advances in Photosynthesis and Respiration 35, Genomics of chloroplasts and mitochondria. Bock, R. and Knoop, V. (eds), Springer, Dordrecht Advances, pp. 103–126.

Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780, doi:10.1093/molbev/mst010 (2013).

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*.

Lau, N. S. Makita, Yuko, Kawashima, Mika, Taylor, Todd D., Kondo, Shinji, Othman, Ahmad Sofiman, Shu-Chien, Alexander Chong , Matsui, Minami, The rubber tree genome shows expansion of gene family associated with rubber biosynthesis. *Sci. Rep.* **6**, 28594 (2016).

Li, B., Lin, F., Huang, P., Guo, W., & Zheng, Y. (2017). Complete Chloroplast Genome Sequence of Decaisnea insignis: Genome Organization, Genomic Resources and Comparative Analysis. *Scientific Reports*, *7*(1), 10073. https://doi.org/10.1038/s41598-017-10409-8

Li, Z., Long, H., Zhang, L., Liu, Z., Cao, H., Shi, M., & Tan, X. (2017). The complete chloroplast genome sequence of tung tree (Vernicia fordii): Organization and phylogenetic relationships with other angiosperms. *Scientific Reports*, *7*(1), 1869.

Lin, Y., Shi, R., Wang, X., & Shen, H. M. (2008). Luteolin, a flavonoid with potential for cancer prevention and therapy. *Current Cancer Drug Targets*, *8*(7), 634-646.

Lohse, M., Drechsel, O., Kahlau, S., & Bock, R. (2013). OrganellarGenomeDRAW--a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research*.

Lowe, T.M. & Chan, P.P. (2016) tRNAscan-SE On-line: Search and Contexture Analysis of Transfer RNA Genes *Nucl. Acids Res.* **44**: W54-57.

Lowe, T.M. & Eddy, S.R. (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.* **25**: 955-964.

Magee AM, Aspinall S, Rice DW, Cusack BP, Se´mon M, Perry AS, et al. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res. 2010;

Mothana, R. A., Gruenert, R., Bednarski, P. J., & Lindequist, U. (2009). Evaluation of the in vitro anticancer, antimicrobial and antioxidant activities of some Yemeni plants used in folk medicine. *Die Pharmazie-An International Journal of Pharmaceutical Sciences*, *64*(4), 260-268.

Okumura S, Sawada M, Park YW, Hayashi T, Shimamura M, Takase H, et al. Transformation of poplar (*Populus alba*) plastids and expression of foreign proteins in tree chloroplasts. Transgenic Res. 2006; 15:637–646. pmid:16952016

Park, S., Jansen, R. K., & Park, S. (2015). Complete plastome sequence of Thalictrum coreanum (Ranunculaceae) and transfer of the rpl32 gene to the nucleus in the ancestor of the subfamily Thalictroideae. *BMC plant biology*, *15*, 40. doi:10.1186/s12870-015-0432-6

Pharm M. (2017) A Study on Anti-cancer activity of *Euphorbia nerifolia*) Milk Hedge) Latex. International Journal of Advances in Science Engineering and Technology, ISSN: 2321-9009,

Raman G, Park S. Analysis of the Complete Chloroplast Genome of a Medicinal Plant, *Dianthus superbus* var. *longicalyncinus*, from a Comparative Genomics Perspective. PLoS ONE 2015;10(10): e0141329 pmid:26513163.

Raman, G.; Park, V.; Kwak, M.; Lee, B.; Park, S. Characterization of the complete chloroplast genome of Arabis stellari and comparisons with related species. PLoS ONE 2017, 12, e0183197.

Rivarola, M., Foster, J. T., Chan, A. P., Williams, A. L., Rice, D. W., Liu, X., … Rabinowicz, P. D. (2011). Castor Bean Organelle genome sequencing and worldwide genetic diversity analysis. *PLoS ONE*, *6*(7), e21743.

Robert M. Waterhouse, Mathieu Seppey, Felipe A. Simão, MoseManni, Panagiotis Ioannidis, GuennadiKlioutchnikov, Evgenia V. Kriventseva, and Evgeny M. Zdobnov

MolBiolEvol, published online Dec 6, 10.1093/molbev/msx319

Roy S, Ueda M, Kadowaki KI, Tsutsumi N. Different status of the gene for ribosomal protein S16 in the chloroplast genome during evolution of the genus Arabidopsis and closely related species. Genes and Genetic Systems 2010; 85:319–326. PMID: 21317544.

Ruhlman T.A. and R.K. Jansen. 2014. The plastid genomes of flowering plants. In: Chloroplast Biotechnology: Methods and Protocols, Maliga, P. (ed), Springer, Humana Press, pp. 3-38.

S. Liu, W. Li, Y. Wu, C. Chen, and J. Lei, "De novo transcriptome assembly in chili pepper (Capsicum frutescens) to identify genes involved in the biosynthesis of capsaicinoids," PLoS ONE, vol. 8, no. 1, Article ID e48156, 2013.

Saeidnia, S., Manayi, A., Gohari, A. R., & Abdollahi, M. (2014). The story of beta-sitosterol-a review. *European Journal of Medicinal Plants*, *4*(5), 590.

Santos, F. A., Frota, J. T., Arruda, B. R., de Melo, T. S., de Castro Brito, G. A., Chaves, M. H., & Rao, V. S. (2012). Antihyperglycemic and hypolipidemic effects of α, β-amyrin, a triterpenoid mixture from Protium heptaphyllum in mice. *Lipids in Health and Disease*, *11*(1), 98.

Saski C, Lee SB, Daniell H, Wood TC, Tomkins J, Kim HG, Jansen RK. 2005. Complete chloroplast genome sequence of Glycine max and comparative analyses with other legume genomes. Plant Mol Biol. 59:309–322.

Steane DA. Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum, *Eucalyptus globulus* (Myrtaceae). DNA Res. 2005; 12:215–220. pmid:16303753

Tangphatsornruang, S., Uthaipaisanwong, P., Sangsrakru, D., Chanprasert, J., Yoocha, T., Jomchai, N., & Tragoonrung, S. (2011). Characterization of the complete chloroplast genome of Hevea brasiliensis reveals genome rearrangement, RNA editing sites and phylogenetic relationships. *Gene*, *475*(2), 104-112.

Ueda M, Fujimoto M, Arimura SI, Murata J, Tsutsumi N, Kadowaki KI. Loss of the *rpl32*gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in *Populus*. Gene. 2007; 402:51–6 pmid:17728076.

Wakasugi, T., Tsudzuki, J., Ito, S., Shibata, M., & Sugiura, M. (n.d.). A physical map and clone bank of the black pine (Pinus thunbergii) chloroplast genome. Plant Molecular Biology Reporter.1994; 12: 227. https://doi.org/10.1007/BF02668746.

Wyman, S. K., Jansen, R. K., & Boore, J. L. (2004). Automatic annotation of organellar

genomes with DOGMA. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bth352.

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829.

Vita

Aldanah Ayidh A Alqahtani has been fascinated by the world of plants since childhood because of the essential connection of botanical knowledge to daily human life. After earning a Bachelor of Science degree in Botany from King Saud University in her native Saudi Arabia in 2012, she taught biology to elementary school students in Riyadh and then was a teaching assistant in biology at Prince Sattam Bin Abdul-Aziz University from 2013-2014. In 2015, she came to the United States to improve her English language skills and completed the University of Dayton's intensive English program as well as Ohio University's intensive English program. In addition, to improve her skills in Plant Biology and gain more experience in laboratory work, she applied to the University of Texas at Austin's master's degree program in Plant Biology. She was accepted into the program in the fall of 2016. Her studies will be focused on generating genomic resources for utilities of toxic plant for medicine purpose.