Investigating the Diversity of Latex Metabolites in Species of the *Euphorbia* Genus

Honors Thesis
Presented to the College of Agriculture and Life Sciences
Cornell University
in Partial Fulfillment of the Requirements for the
Biological Sciences Honors Program
Plant Biology Section

by
Se Jin Park
May 2021

Faculty Mentor: Dr. Gaurav Moghe

TABLE OF CONTENTS

ABSTRACT

The Euphorbiaceae family is one of the largest angiosperm families with its largest genus *Euphorbia*, containing about 2000 species. This genus is characterized by the presence of laticifers and latex. Latex is a milky fluid with diverse secondary metabolites but especially enriched with terpenes. Many of these metabolites, especially diverse terpenes, are used by the Euphorbias as a defense mechanism against herbivores. Although *Euphorbia* is a large genus, it is relatively understudied at the metabolomic and genomic level compared to other Euphorbiaceae genera. The goal of this research was to establish a foundation for further investigation into latex, laticifers and their contribution to species diversity in the *Euphorbia* genus. Towards this goal, first, I performed a metabolomics study of 18 *Euphorbia* species. Next, two model Euphorbias – *Euphorbia peplus* and *Euphorbia lathyris* – were selected for further analysis based on previous biochemical studies. We confirmed their identity and estimated their genome size using *maturase K* (*matK*) and flow cytometry analysis, respectively. Further, a latex metabolite extraction protocol was standardized and used for assessing the diversity of metabolites from the latex of the two *Euphorbia* model species using a machine learning approach. The metabolite networking analysis suggested that the *Euphorbia* latex has a diverse network of secondary metabolites, from which we identified 13 unique terpenes. Finally, to understand the genetic causes of the metabolic diversity, *de novo* assembled transcriptomes were used to identify terpene synthase genes. Phylogenetic analysis revealed that the Euphorbiaceae family and the *Euphorbia* genus have conserved terpene synthase subfamilies in relation to other plant families that were studied. The results indicate that the *Euphorbia* latex is highly diversified while having relatively conserved proteins that are responsible for producing the latex metabolic diversity.

Key Words: Euphorbiaceae, *Euphorbia*, laticifer, latex, *maturase K* (*matK*), flow cytometry, LC-MS, MSDIAL, metabolites, terpenes, terpene synthases (TPSs), phylogeny

1. INTRODUCTION

1.1 The relation between angiosperm diversity and secondary metabolite diversity

Angiosperms, also known as flowering plants, house ~90% of all land plant species, covering ~300,000 species (Wiens and Hernandez, 2019; Christenhusz and Byng, 2016). There are many potential factors such as ecological factors, physiological and morphological differences, differences in pollination and herbivore pressure, and metabolite diversity that are responsible for such a high diversification rate (Kubitzki and Gottlieb, 1984; Crepet, 1984). Above all, diverse secondary metabolites produced from the primary metabolite precursors (sugars, amino acids, nucleotides, and fatty acids) are crucial for plant survival in various ecological conditions. There are over one million metabolites produced by the entire plant kingdom (Maeda, 2019). Plant secondary metabolites not only serve as a defense mechanism but also as an attractant for various pollinators. This metabolic diversity across the plant kingdom is a result of frequent gene duplications and enzyme promiscuity that occurs at the genetic level, and which is selected via various ecological processes (Wink, 2020). One model to explain the increase in angiosperm diversity is the “coevolutionary” hypothesis between plant secondary metabolite diversity and herbivory (Foisy et al., 2019). As per this hypothesis, a constant “arms race” between plants and their herbivores leads to increases in metabolic diversity as well as promotes a higher degree of speciation (Ehrlich and Raven, 1964). Many specialized metabolites are also acquired in various plant lineages through convergent evolutions (Pichersky and Lewinsohn, 2011).

Within the angiosperm clade, the Euphorbiaceae family is one of the most diverse plant families. Despite its large size, the Euphorbiaceae family seems to be monophyletic (Webster, 1994). Many plants in the family produce various chemical compounds/metabolites, which are often poisonous but also useful (Ernst et al., 2019). Such diverse chemical composition makes the Euphorbiaceae family a great model for metabolomics study.

1.2 The Euphorbiaceae is one of the largest families in the plant kingdom, but the causes for its diversification are not well-understood

Within the angiosperm group, the Euphorbiaceae family is the fifth largest family with three subfamilies (Acalyphoideae, Crotonoideae, and Euphorbiodeae), 37 tribes, 300 genera, and 7,500 species (Armbruster et al., 1997). This diverse group comes in various forms such as succulents, herbs, shrubs, and trees. Euphorbiaceae species are found in habitats ranging from tropical areas to temperate regions on all continents except Antarctica (Perry, 1943; Ernst et al., 2016). Many Euphorbiaceae species have various economic benefits. *Euphorbia peplus* can be used to treat skin cancer and the precancerous skin condition actinic keratosis (Ernst et al., 2015), and *Euphorbia lathyris* is used as an active ingredient of Picato® medicine for treating actinic keratosis (Salehi et al., 2019). *Croton celtidifolius* – belonging to the second largest genus in the family – is used for treating inflammatory diseases, leukemia, ulcer, and rheumatism, and *Croton arboreous* is used for treating respiratory ailments (Salatino et al., 2007). In addition, *Hevea brasiliensis* is an important source of rubber (Van Parijs et al., 1991), *Manihot esculenta* is one of the largest sources of carbohydrates in the tropics (Dufour, 1988), and *Jatropha curcas* seed oil can be used for the biofuel production (Maghuly et al., 2015). As horticultural decorations, *Euphorbia pulcherrima* and *Ricinus communis* are often used (Maghuly et al., 2015).

Two of the largest genera in the Euphorbiaceae family are *Euphorbia* with ~2000 species and *Croton* with 1300 species (Salatino et al., 2007). There are four subgenera, which include *Chamaesyce, Esula, Euphorbia,* and *Rhizanthium* within the *Euphorbia* genus (Bruyns et al., 2006). The *Euphorbia* genus is characterized by latex production (**Section 1.3**) and cyathium, which is a reduced cluster of flowers. This special structure is only present in the *Euphorbia* genus. The main feature of cyathium is the floral envelope or involucre that surrounds each group of flowers (Prenner and Rudall, 2007). The cyathium is considered a potential cause of diversification of the *Euphorbia* genus because this inflorescence structure potentially allowed this genus to shift away from wind pollination to insect pollination (Horn et al., 2012).

In addition, the *Euphorbia* genus is unique in that the plants within this group perform all three types of photosynthesis: C3, C4, and CAM (Webster et al., 1975). C3 plants have a normal photosynthetic mechanism and do not have photosynthetic adaptations, leading them to photorespiration. C4 plants have physically separated structures (mesophyll and bundle sheath cells) that perform light-dependent reactions and the Calvin cycle respectively. On the other hand, plants that perform CAM photosynthesis live in arid conditions and have temporal separation, which means that CAM plants allow CO_2 to enter at night and utilize that adaptation to improve photosynthetic efficiency. The *Euphorbia* genus not only has different ways of performing photosynthesis but also grows in diverse ecological habitats. Herbaceous Euphorbias are found in temperate zones worldwide while succulent plants are found in southern/eastern Africa and Madagascar/tropical Asia and the Americas (Webster, 1994). Thus, it has been proposed that the C4 and CAM adaptations helped contribute to the high species diversity of the *Euphorbia* genus (Horn et al., 2014).

1.3 Laticifers are specialized cells that produce latex

In addition to cyathium and the different photosynthetic mechanisms, latex production is also one of the defining features of the *Euphorbia* genus. The broader study of which this thesis project is part of seeks to identify the role of latex/defense metabolite diversity in increasing the species diversity in Euphorbiaceae.

Latex is a milky fluid (either white, colorless, or red) containing various proteins and secondary metabolites such as terpenoids, alkaloids, cardenolides, glycosides etc. along with various proteins and chemicals (Konno, 2011). It is hypothesized that the latex is used for removing waste metabolites, dealing with damaged plant tissue, and defending against herbivores (Konno, 2011). Latex is the cytoplasm of a specialized cell called the laticifer, which can span the entire plant body (Ramos et al., 2020). Approximately 40 angiosperm families, which includes more than 20,000 species, have laticifers that produce latex (Lewinsohn, 1991).

There are two different types of laticifers: non-articulated laticifers and articulated laticifers. Non-articulated laticifers are formed from few initial cells and eventually branch without cell division and form multicellular tube cells. These laticifers have tree-like shapes but do not form loops (net-like structures) as laticifer branches do not merge. Angiosperm families that have non-articulated laticifers include Moraceae, Caricaceae, and Euphorbiaceae (Konno, 2011). Most Euphorbiaceae family members including the *Euphorbia* genus have non-articulated laticifers. However, the subfamily Crotonoideae contains articulated laticifers (Hagel et al., 2008). Since not all Euphorbiaceae genera or species produce latex, the *Euphorbia* genus is unique in that all species produce latex and have non-articulated laticifers (Rudall, 1994). This unifying characteristic within this large genus makes it suitable for studying unique qualities within the *Euphorbia* genus while understanding how this genus became so diverse. Some other angiosperm families including

Convolvulaceae, Anacardiaceae, and some Caricaceae have articulated laticifers. These are developed from the longitudinal chain of cells that form tube structures because the cell walls that separate the chain of cells vanish (Konno, 2011). Latex found in various angiosperm families contains diverse secondary metabolites that serve as defense mechanisms.

1.4 Diverse secondary metabolites found in latex

Previously, it was believed that the plants produce latex to excrete waste metabolites and reserve water. However, with further research, currently, there is greater evidence that the plants produce latex for defense mechanisms against herbivores, insects, and pathogens (Hua et al., 2017). Latex contains various types of secondary metabolites such as terpenoids (diterpenes, triterpenes, etc.), glycoside, alkaloids, flavonoids, cardenolides, lignan, coumarins, and phenols that are toxic against herbivores, insects, and pathogens (Hua et al., 2017; Sharma et al., 2014; Ramos et al., 2019). Of all secondary metabolites, terpenes and terpene derivatives are the main components of *Euphorbia* latex (Nemethy et al., 1983), and play important roles in *Euphorbia* herbivore defense (Hua et al., 2017).

1.5 Latex as a defense mechanism led to plant diversification as addressed by the coevolution hypothesis

Plant evolution and diversification rates are related to defense mechanisms, one example being latex production (Farrell et al., 1991). The relationship between the evolution of laticifer/latex and plant family speciation is described by the escape and radiate coevolution hypothesis (Foisy et al., 2019). This hypothesis illustrates how the evolution of the plant defense mechanism is related to the coevolution between herbivores and plants. It is not clear if latex

production is directly correlated to plant diversification. Plants release latex from their leaves and stems when attacked by herbivores and insects. Latex can be used as a defense mechanism because it is toxic. Produced by ~10% of angiosperm species, latex evolved about 40 independent times (Foisy et al., 2019). However, little research has been conducted to determine if there is a correlation between the diversity of these secondary metabolites and species diversity. One recent study identified a poor correlation between species diversity and the presence of laticifers and resin canals across plants (Foisy et al., 2019), however, lack of a group-level resolution. Such potential correlation has not been robustly studied for the Euphorbiaceae family (Farrell et al., 1991).

Latex not only contributes to plant family diversity but also has important applications in medicine production and rubber industry. Depending on the dose, latex can be either toxic or medicinal (Basak et al., 2009). The *Euphorbia* latex has various usages and benefits. For instance, anti-feedant effects of *E. peplus* latex towards *Helicoverpa armigera* have been studied (Hua et al., 2017). Furthermore, *Euphorbia tirucalli* latex has been used as Brazilian anticancer medicine (de Souza et al., 2019). Plant latex can also have economic benefits such as the production of rubber from *Euphorbia characias* latex (Spanò et al., 2012). In addition, *Euphorbia amygdaloides* latex protease is used for cheese production (Demir et al., 2005). Such diverse usages make latex an interesting topic to study both in terms of plant diversification and potential medicinal impacts.

Euphorbia latex contains various secondary metabolites that are beneficial for plants. These metabolites, especially terpenes, potentially allowed the *Euphorbia* genus to defend against herbivores and become diverse. Also, latex serves as a great economic resource due to its diverse secondary metabolites. In order to understand and learn about *Euphorbia* latex composition and diversity, latex samples from *Euphorbia* plants were extracted and studied. We further studied

terpene synthases (TPSs) from two model species to understand the genetic mechanisms of latex terpene diversification.

1.6 Project overview

In this project, a model *Euphorbia* species was selected, the diversity of latex secondary metabolites was characterized, and a foundation for enzyme characterization was established by accomplishing five major objectives. First, a latex extraction protocol was developed to study *Euphorbia* latex to understand the secondary metabolite diversity. Next, using this protocol, secondary latex metabolite diversity of *E. peplus* and 17 other *Euphorbia* species was studied using the MS1 results from the liquid chromatography-mass spectrometry (LC-MS) and computational data analysis methods. Mass spectrometry is used to measure the mass to charge ratio (m/z) of the molecules (metabolites) that are present in a sample (latex). This analysis helps with quantifying and identifying the type of metabolites. After performing the metabolomics analysis, a model *Euphorbia* species, *E. peplus*, was selected based on its relatively shorter life cycle compared to other potential *Euphorbia* model species, successful germination, and seed production (Mergner et al., 2020). *E. peplus* identity was assessed and verified by the *maturase K* (*matK*) analysis and the genome size was measured using flow cytometry analysis (Arumuganathan and Earle, 1991b). The *matK* is a chloroplast gene that is suggested to serve as a barcode for land plants (Yu et al., 2011). *E. peplus* and *E. lathyris* latex were further assessed using LC-MS/MS to understand the secondary metabolite diversity and metabolites, especially terpenes, that are unique for these two species based on the differential metabolite accumulation analysis. Finally, after identifying the terpene diversity, the phylogenetic study was performed on *E. peplus* and TPS diversity was

studied. The methodological and biological resources and knowledge developed in this project will allow researchers to assess the molecular causes of diversification in the *Euphorbia* genus.

2. MATERIALS & METHODS

2.1 Germinating and transplanting *Euphorbia* plants

Seed packets for *E. peplus*, *E. lathyris*, *Euphorbia prostrata*, and *Solanum lycopersicum* (**Section 2.3)** were provided and these seeds were germinated using Moghe Lab's previously established trisodium phosphate (TSP) sterilization and seed germination protocol. Fifteen other *Euphorbia* species along with two outgroup species (**Section 2.2)** were also germinated using this protocol. Seeds were incubated in 10% TSP for 10 minutes. *Solanum lycopersicum* was not used for **Section 2.2**. Sterilized seeds were transferred to the prepared Petri dishes and 1 mL of sterile deionized water with 200 uM Gibberellic acid 3 (GA3) was added to enhance the germination rate. Petri dishes were monitored daily, and 1 mL of water was added when the filter paper dried out. Germinated seeds were then transplanted into Cornell mix (peat-lite mixes) (Boodley and Sheldrake, 1972). The *Euphorbia* plants require full sun exposure, so these plants were kept in the growth chamber with long day conditions (16h light: 8h dark) and watered regularly. Following this protocol, a total of 35 *E. peplus*, seven *E. lathyris*, one *E.prostrata*, and 15 *Solanum lycopersicum* were germinated.

2.2 Latex extraction and *Euphorbia* latex metabolomic diversity analysis

Developing the latex extraction protocol

Latex has both polar and non-polar components, requiring the use of a mixture of polar and non-polar solvents (Verma, 2013). Some commonly used solvents such as methanol, dimethyl

sulfoxide (DMSO), isopropyl alcohol (IPA), H_2O, acetonitrile (ACN), isopropanol, and hexane were selected based on their polarities and previous experiments done by the Moghe Lab. Water with a relative polarity of 1 was used as the main solvent. Another solvent that was nonpolar enough to dissolve latex but polar enough to mix with water was selected. Plausible solvents were DMSO, methanol, ACN, and IPA. Seven different protocols were tested for the latex extraction (**Section 3.1**).

Optimal latex extraction protocol

The weight of a capillary tube with 1.5 mm outer diameter in the 15 mL Falcon tube was measured. This weight was tared to only account for the collected latex weight. The Falcon tube was labeled with corresponding species. After setting up collection tubes, small incisions were made in the stems, branches, or leaves of plants using a new razor blade for each sample. Using the capillary tube, latex oozing out of the incision was collected and then the capillary tube was placed in the Falcon tube and the weight of the latex was measured. Each Falcon tube was labeled with the amount of extraction solvent based on the weight of the latex.

Using forceps, the capillary tube was pulled out. Using a micropipette, the latex was aspirated into the Falcon tube. 200 uL of extraction solvent (100 uL of ACN and H_2O) per 10 mg of latex was directly added into the capillary tube to aspirate the remaining latex. The capillary tubes were disposed of. The sample was vortexed for 1 min and mixed using the shaker for 20 mins. The Falcon tube was centrifuged for 10 mins at 14,000 x g, 22^{o}C. The supernatant was transferred into an HPLC vial.

Running LC-MS

LC-MS analysis was performed on a ThermoScientific Dionex Ultimate 3000 HPLC equipped with an autosampler coupled to a ThermoScientific Q-Exactive HF Orbitrap mass spectrometer using solvent A (water + 0.1% FA) and solvent B (ACN) at a flow rate of 0.6 ml/min using a gradient. Several LC method parameters were set (**Table 1**). Metabolites were detected in positive and negative ionization modes. MS2 (second stage mass spectrometry) was not performed. The total scan range used was 66 to 990 m/z. For MS1 (first stage mass spectrometry) following parameters were used: resolution: 140,000; AGC target: 3e6; maximum IT 110 ms.

Table 1. LC method parameters

Time (min)	Solvent
0 – 5.6	PumpModule.Pump.%B.Value: 0.0 [%]
5.6 – 6.61	PumpModule.Pump.%B.Value: 100.0 [%]
6.61 – 7.510	PumpModule.Pump.%B.Value: 0.0 [%]
All at curve 5	

Creating projects using the MS-DIAL program and modifying with Excel

After the MS1 data of the extracted latex samples were processed using the LC-MS instrument (completed by Alexandra Bennett), secondary metabolite diversity analysis was performed using MS-DIAL (ver. 4.48) (Tsugawa et al., 2015). For this analysis, both the positive and negative ion modes were used. First, obtained raw files from the LC-MS were converted to Analysis Based Files (ABFs) using the Reifycs ABF Converter, which were selected to create an MS-DIAL project. To set up for the new positive mode MS-DIAL project, soft ionization, conventional LC/MS or data dependent MS/MS, centroid data (for both data type – MS1 and MS/MS), positive ion mode, and metabolomics were selected.

After creating a new project, an internal standard (P4HB) was identified using the molecular weight and retention time. The project was normalized to the internal standard within MS-DIAL. The blank filter option was added. Lastly, the alignment result – including five blank samples, 18 *Euphorbia* samples, and two outgroup samples (**Supplement Part 1**) – was exported along with an mgf file. When exporting, normalized data matrix, GNPS export, and filtered by blank peaks, were selected. The normalized project was used to perform principal component analysis (PCA). For the negative ion mode, most settings were kept the same with some changes. Different adducts were selected, which included [M-H]- and [M+FA-H]-. The internal standard was again identified in negative mode.

Different filters were applied to the exported Excel file in order to detect metabolites that are present in plant samples. The raw threshold (not normalized) of 10,000 was applied for the species of interest – if a metabolite has a value > 10,000, it is confidently present in the species of interest. Additionally, a filter of 1,000 was applied for the species not in interest – if a metabolite has a value < 1,000, it is confidently absent in the species. For positive and negative ion mode, the filter greater than or equal to 10,000 was applied to all plant samples, while the blank sample was left alone. This step was used to identify the total number of metabolites that are present in all plant samples (*Euphorbias* and outgroup plant samples) in each ion mode. For the next analysis, the filter greater than or equal to 10,000 was only applied to one plant sample to determine the total number of metabolites that are present in each plant sample. This process was repeated for all plant samples. Lastly, the filter of greater than or equal to 10,000 was applied to a plant sample of interest and then a filter of less than or equal to 1,000 was applied to all other plant samples in order to find unique metabolites for each plant sample. This process was also repeated for all plant

samples. The unique metabolite ID numbers found for *E. peplus* and *E. lathyris* were identified on the MS-DIAL program to evaluate the peak shapes.

Generating PCA and Shannon Entropy estimates

Principal component analysis (PCA) using MS-DIAL was performed to understand the sample clustering based on metabolomic abundance. The default setting was used for the PCA analysis and the labels were removed. For Shannon Entropy analysis, raw data matrix area files for both positive and negative ion modes were exported as an excel file using MS-DIAL. An in-house script (part of Moghe Lab's current research and is not published) was used to calculate the diversity of metabolites as measured by Shannon Entropy for each sample (each species). The Shannon Entropy results were visualized using R software (ver.4.0.2) (R Studio Team, 2020) and ggplot2 (ver.33.2) (Wickham, 2009). These steps for creating Shannon Entropy graphs were completed by Elizabeth Mahood.

2.3 Identifying and selecting model *Euphorbia* species

Maturase K (*matK*) analysis

In order to verify the model *Euphorbia* species for this project, *matK* sequences of *E. peplus* and *E. lathyris* were assessed. Two pairs of forward and reverse primers were selected based on the *matK* sequence analysis. Primers *trnK* 570F and 1710R (Samuel et al., 2005) along with *matK*_F1 and *matK*_R1 designed using the Geneious Prime software (ver. 2021.0.1) (Kearse et al., 2012) were used for amplification and Sanger sequencing of the *matK* sequences of *E. peplus* and *E. lathyris* (**Table 2**).

Table 2. Maturase K (*matK*) forward (F) and reverse (R) primers

Primer Name	Sequence
trnK 570F (Samuel et al., 2005)	5'-TCC AAA ATC AAA AGA GCG ATT GG-3'
1710R (Samuel et al., 2005)	5'-GCT TGC ATT TTT CAT TGC ACA CG-3'
*matK*_F1	5'-CCC CAT CCA TCT CGA AAA ATT GG-3'
*matK*_R1	5'-ATA CGC GCA AAT TGG TCG AT-3'

The DNA extraction was performed using the E.Z.N.A. SP Plant DNA Kit (Omega Bio-tek) from fresh plant samples. PCR was performed based on Moghe lab's previously established protocol using the NEB Q5 PCR kit. The DNA sample was amplified based on these thermal cycling conditions: 98 ^{o}C for 1 min, 98 ^{o}C for 10 sec, 53 ^{o}C for 1 min, 72 ^{o}C for 1 min, followed by 34 cycles of 98 ^{o}C for 10 sec, and 72 ^{o}C for 5 min with a final temperature set up of 12 ^{o}C. Gel electrophoresis was performed using 1% agarose gel, and samples with confirmed amplicon bands were sent for sequencing at Cornell's Biotechnology Resource Center. The sequences thus obtained were compared with the *matK* sequences from the National Center for Biotechnology Information (NCBI)'s GenBank database

Flow cytometry analysis overview

Flow cytometry analysis was performed to measure the genome sizes of *E. peplus* and *E. lathyris*. Two different published protocols with some modifications were used to perform flow cytometry analysis – one using 4', 6-diamidino-2-phenylindole (DAPI) as a fluorochrome (Pellicer and Leitch, 2014) and another using propidium iodide (PI) as a fluorochrome instead of DAPI (Arumuganathan and Earle, 1991b). However, we realized later that the DAPI protocol was not suitable for this project because DAPI is an AT base-specific fluorochrome and thus was not

suitable for estimating the genome size (Pellicer and Leitch, 2014). Therefore, the flow cytometry protocol using PI was selected (Arumuganathan and Earle, 1991a).

Flow cytometry solvent preparation

For this experiment, several stock solvents ($MgSO_4$ buffer, Triton X-100 stock (10% w/v), and PI stock (5 mg/mL)) were prepared along with the provided RNAse solution (DNAse free) and Chicken Red Blood Cells (CRBCs). This protocol had another stock solvent (Alsever's solution) to dilute the CRBC solution, but the purchased CRBC solution had an ideal concentration of ~10^7 CRBC/mL. Thus, Alsever's solution was not prepared. Using these stock solutions, three solutions for the flow cytometry analysis (**Table 3**).

Table 3. Solutions A, B, and C composition

Solution	Composition
Solution A (15 mL enough for 12 samples)	14.3 mL $MgSO_4$ buffer 15 mg dithiothreitol (Sigma, D-0632) 300 uL PI stock 275 uL Triton X-100 stock
Solution B (~3 mL enough for 12 samples)	3 mL Solution A 7.5 uL RNAse 3.0 uL CRBC
Solution C	3 mL Solution A 7.5 uL RNAse
All solvents were stored in the refrigerator at 4 °C	

Flow cytometry sample preparation

Genomic DNA was produced as previously described (Arumuganathan and Earle, 1991b). First, 50 mg of leaf tissue was placed on a glass Petri dish located on the ice bucket. 1 mL of Solution A was added, and the leaf tissue was finely chopped with a razor blade for 7 – 10 mins.

After gently shaking the Petri dish, the homogenate was filtered through a 30 – 42 um filter into the 25 mL falcon tube. The filtered sample was aliquoted into 1.5 mL Eppendorf tubes. These samples were centrifuged at 10,062 g for 15 – 20 secs. The supernatant was discarded, and the pellet was resuspended in 200 uL of Solution C. This step is a modification of the original protocol because the CRBC peak overlapped with the plant genomic DNA peaks and CRBC was not suitable as an internal standard. Lastly, samples were incubated for 15 mins at 37°C.

Setting up and running the flow cytometry machine

Attune NxT Flow Cytometer (Thermo Fisher Scientific) in the Biotechnology Resource Center was used to perform the flow cytometry analysis. Using this machine, voltages were controlled to an optimal setting based upon graph centering, and additional graphs of interest were added (**Figure 1**). In addition, the gates were set around the interested population. The prepared plant sample along with incubated CRBC sample (Solution B) was used for the flow cytometry analysis. After collecting graphs of interest, the FlowJo program was used to recreate those graphs. Lastly, genome sizes were calculated using the equation: $nuclear\ DNA\ amount = \left[\frac{mean\ position\ of\ plant\ nuclear\ peak}{mean\ positon\ of\ CRBC\ nuclear\ peak}\right] \times 2.33\ pg$. CRBC nuclei have a genome size of 2.33 pg/2C (Arumuganathan and Earle, 1991a).

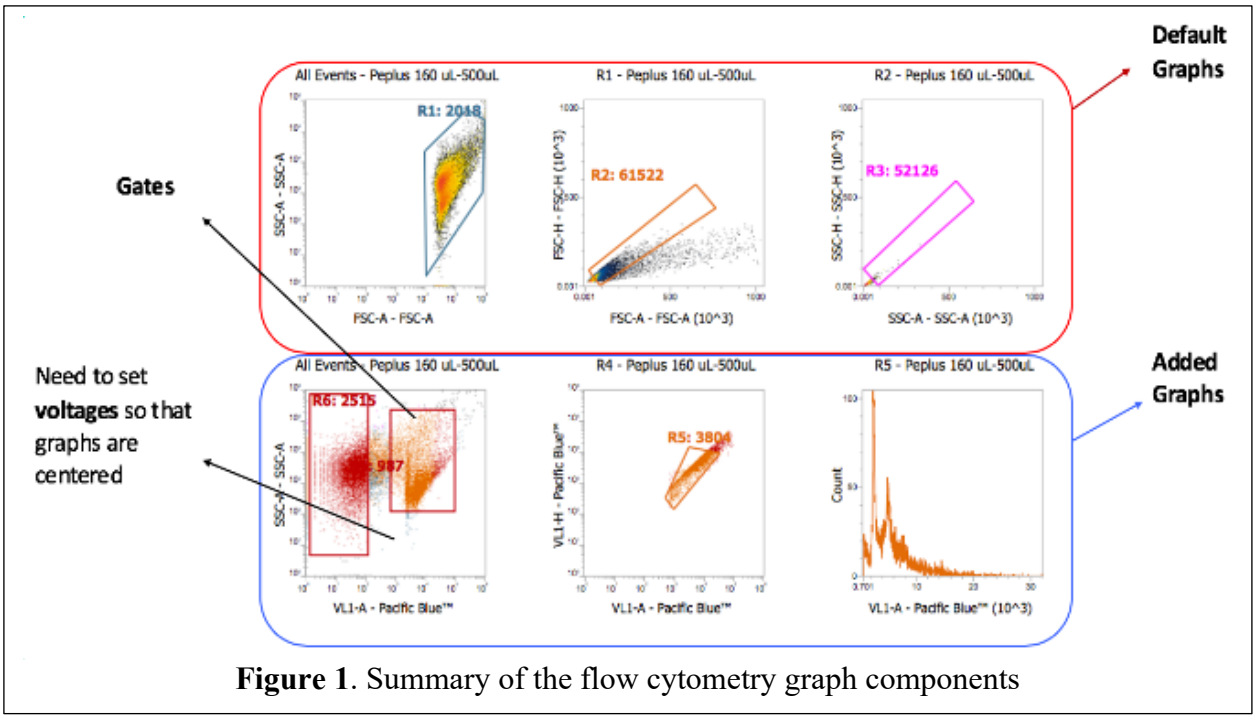

Figure 1. Summary of the flow cytometry graph components

2.4 Metabolomic analysis of the *E. peplus* and *E. lathyris* latex secondary metabolites

LC-MS analysis was performed on a ThermoScientific Dionex Ultimate 3000 HPLC equipped with an autosampler coupled to a ThermoScientific Q-Exactive HF Orbitrap mass spectrometer using solvent A (water + 10 mM ammonium formate, pH 3.2) and solvent B (acetonitrile+ 0.1 %(v/v) formic acid) at a flow rate of 0.6 ml/min using a gradient. Several LC method parameters were set (**Table 4**).

Table 4. LC method parameters

Time (min)	Solvent
0 – 0.5	5% solvent B
0.5 – 2	Gradient from 5% solvent B to 30% solvent B (gradient curve 5)
2 – 18	Gradient from 30% to 95% of solvent B (gradient curve 5)
18 – 18.5	Solvent B kept at 95%
Afterward	Column was re-equilibrated from 18.5 min to 20 min with 5% solvent B

Metabolites were detected using either full MS1 or full MS1 coupled to a data-dependent MS2 method, extracting the top 3 most abundant ions (Full-MS/ddMS2) in positive ionization mode. The total scan range used for MS1 was 67 to 1,000 m/z. For MS1 following parameters were used: resolution: 70,000; AGC target: 3e6; maximum IT 200 ms. For MS2 the parameters were: resolution: 35,000; AGC target: 3e5; maximum IT: 100 ms (Provided by Lars Kruse).

First, obtained raw files from the LC-MS were converted to mzML files using the ProteoWizard MSConvertGUI (ver. 3.0.11781) (Adusumilli and Mallick, 2017). These mzML files were used to create an MS-DIAL project. To set up for the new MS-DIAL project, soft ionization, conventional LC/MS or data dependent MS/MS, Centroid data (for both data types – MS1 and MS/MS), positive ion mode, and metabolomics were selected. This alignment file includes blanks along with three *E. lathyris* and six *E. peplus* samples were included (**Supplement Part 2**).

After creating a new MS-DIAL project, an internal standard (telmisartan) was identified using the molecular weight and retention time. An mgf and a peak height alignment file were exported using the MS-DIAL software (ver. 4.48) (Tsugawa et al., 2015). That mgf was then used by Canopus and Sirius (ver. 4.6.1) for compound annotation. Within Sirius, all the parameter values were set as default except for two values. The “Instrument” value was set to Orbitrap and the “Candidates” value was set to 3. Sirius predicts three compounds for each mgf entry and Canopus assigns the class to each of those three compounds. This part was completed by the Moghe Lab’s graduate student Elizabeth Mahood. The compound that is most likely based on the prediction by Sirius was used.

To perform the molecular networking study, MS-FINDER (ver. 3.50) program was used to generate files. First, an mgf file exported from the MS-DIAL was separated into MSP files.

These MSP files were imported. For the setting, mass tolerance [Da] was set to 0.01, relative abundance cut off [%] was set to 5, MS/MS similarity cut off [%] was set to 75, retention time tolerance [min] was set to 100, add metabolite ontology bioreactions was selected, and retention time definition for 'near' was set to 0.5. Next, under the export tab, export nodes and edges for the molecular spectrum network were selected. An in-house script was used to annotate each node with Canopus identified compound class. Queries that are identified by Canopus as prenol lipids and steroids and steroid derivatives were labeled as terpenes. Cytoscape was used to visualize the node and modified edge file using the Prefuse Force Directed Layout, which is the default format. A second network was created only using the metabolites that survived the blank filter and adduct filter. Each metabolite's superclass was identified for this molecular network (this file was created by Elizabeth Mahood).

In addition to the molecular networking map, volcano plots were created. First, the exported normalized peak height file was modified using Excel 2016. All sample values (blank and plant samples) were divided by the internal standard value. The blank filter was applied for any values that had a blank average greater than or equal to 20% of the sample max were removed. Using this file, a bar graph of secondary metabolite superclasses and volcano plots for terpenes and all secondary metabolites were created using Excel. For the volcano plots, the p-values were corrected with Benjamini Hochberg correction method.

2.5 Phylogenetic analysis of terpene synthases

Generating TPS sequences

First, the proteomes of *Hevea brasiliensis*, *Jatropha curcas*, *Manihot esculenta*, and *Ricinus communis* were downloaded as FASTA files from the NCBI website. As there is no

sequenced genome for *E. peplus* and *E. lathyris*. *De novo* assembled transcriptome FASTA files for *E. peplus* and *E. lathyris* were used instead (transcriptomes were created by the Moghe Lab's former visiting scientist Dr. Kai Fan). These transcriptomes only include the longest isoform.

Using the Pfam website, profile Hidden Markov Models (HMM) files for the C-terminal and N-terminal domains of the TPS were downloaded and used to perform hmmsearch using the command-line version of the HMMER (ver. 3.3.1) program. Hmmsearch uses the HMM file to search for sequences that are similar. The general command line used for this analysis is listed under the supplement section (**Supplement Part 3**). At the end of the hmmsearch analysis, output files for C-terminal and N-terminal for each species were analyzed to identify sequences containing both the C-terminal and N-terminal domains. Such sequences were annotated as high-confidence TPS.

Developing phylogenetic trees and performing analysis

A phylogenetic tree of TPS from all Euphorbiaceae samples along with *Arabidopsis*, *Oryza*, *Physcomitrella*, *Populus*, *Selaginella*, *Sorghum*, and *Vitis* was generated (Chen et al., 2011). Specifically, the TPS protein sequences were aligned using MAFFT (ver.7.453-with-extensions) (Katoh, 2002) in Geneious Prime (ver. 2020.1.1) using default parameters, and the alignment was provided as input to IQ-TREE (ver.1.6.10) (Nguyen et al., 2015), which was run with model selection (ModelFinder (Kalyaanamoorthy et al., 2017)) and 1000 ultrafast bootstrap replicates (Minh et al., 2013) using following parameters: *-st AA -nt AUTO -ntmax 10 -bb 1000 -m TEST*. In each analysis, the optimal tree was obtained using the TVM+F+G4 model, identified based on Model Finder results. The TPS groups were assigned to each predicted TPS in the *Euphorbia* species by comparing the published phylogenetic tree and the sequences used for the published

papers to the sequences that were created for this project. TPS subfamilies were assigned to the accession number from the supplementary table (Chen et al., 2011) based on published articles (Aubourg et al., 2002; Lin et al., 2017; Jiang et al., 2019; Martin et al., 2010; Jones et al., 2011; Hillwig et al., 2011; Dudareva et al., 2003; Yu et al., 2020). After assigning the correct subfamilies, protein sequences were obtained in a FASTA format using various genome browsers: TAIR (Rhee, 2003), Phytozome (Goodstein et al., 2012), and Michigan State University Rice Genome (Ouyang et al., 2007). Using the Geneious program, custom BLAST sets were created using the TPS sequences from the hmmsearch analysis. Sample TPS sequences from the non-Euphorbiaceae species listed above from the supplementary table (Chen et al., 2011) were compared with the hmmsearch sequences.

3. RESULTS

3.1 Comparing and selecting a latex extraction protocol

Given that understanding the latex metabolic diversity was the primary goal of this research, we first tested seven different protocols to optimize latex metabolite extraction. Although these extraction protocols were developed using the latex from a Moraceae species, the optimal protocol was selected based on the results from the latex of Euphorbiaceae species, using the number of metabolites and the number of steps required by each protocol as selection criteria. Based on UHPLC-MS/MS analysis, protocol 7 (**Table 5**) i.e., extraction in 1:1 H_2O: acetonitrile mixture resulted in a similar number of metabolites as protocol 6 with fewer steps, allowing protocol 7 to be cost and time-efficient. The optimal latex extraction protocol was further used to estimate latex metabolite diversity of the 18 *Euphorbia* species.

Table 5. Summary of seven extraction protocols and outcomes

Protocol	Summary	Outcomes
1	First extraction in 100% H_2O Second extraction in excess 100% DMSO	Latex was dissolved, but the final solution was too diluted. This protocol wasn't further used for comparison.
2	Uses dried latex The solvent system was not selected as latex did not dry for further experiment.	Dried latex was used, but it was inefficient as latex would not dry quickly (Züst et al., 2019).
3	First extraction in 1:1 MTBE: MeOH Second extraction in 1:1 H_2O: MeOH (Salem et al., 2017)	This protocol was based on the previously published biphasic method, but the two immiscible liquid phases did not separate (Salem et al., 2017).
4	First extraction in 100% H_2O Dissolve pellet in 100% ACN	This protocol produced one final sample. First comparison of protocols 4, 5, and 6: • *Euphorbia characis wulfenii*: 2125 metabolites • *Euphorbia* 'Rose's Surprise': 1492 metabolites
5	Extraction in 1:1:1 IPA: ACN: H_2O	This protocol produced two final samples. First comparison of protocols 4, 5, and 6: • *Euphorbia characis wulfenii*: 3212 metabolites • *Euphorbia* 'Rose's Surprise': 1215 metabolites
6	First extraction in 100% H_2O Dissolve pellet in 100% ACN Combine first extraction with dissolved pellet	This protocol produced one final sample. First comparison of protocols 4, 5, and 6: • *Euphorbia characis wulfenii*: 2946 metabolites • *Euphorbia* 'Rose's Surprise': 1037 metabolites

		Second comparison between protocols 6 and 7: • *Euphorbia guentheri*: 2357 metabolites • *Euphorbia meloformis*: 2398 metabolites
7	Extraction in 1:1 H_2O: ACN mixture	This protocol produced one final sample and required the least steps. • *Euphorbia guentheri*: 2194 metabolites • *Euphorbia meloformis*: 2728 metabolites

3.2 MS1 scans across 18 different *Euphorbia* species reveal substantial metabolic diversity in *Euphorbia* latex

Metabolomic analysis was performed using the optimal extraction protocol to understand the latex secondary metabolite diversity within the *Euphorbia* genus (**Figure 2**).

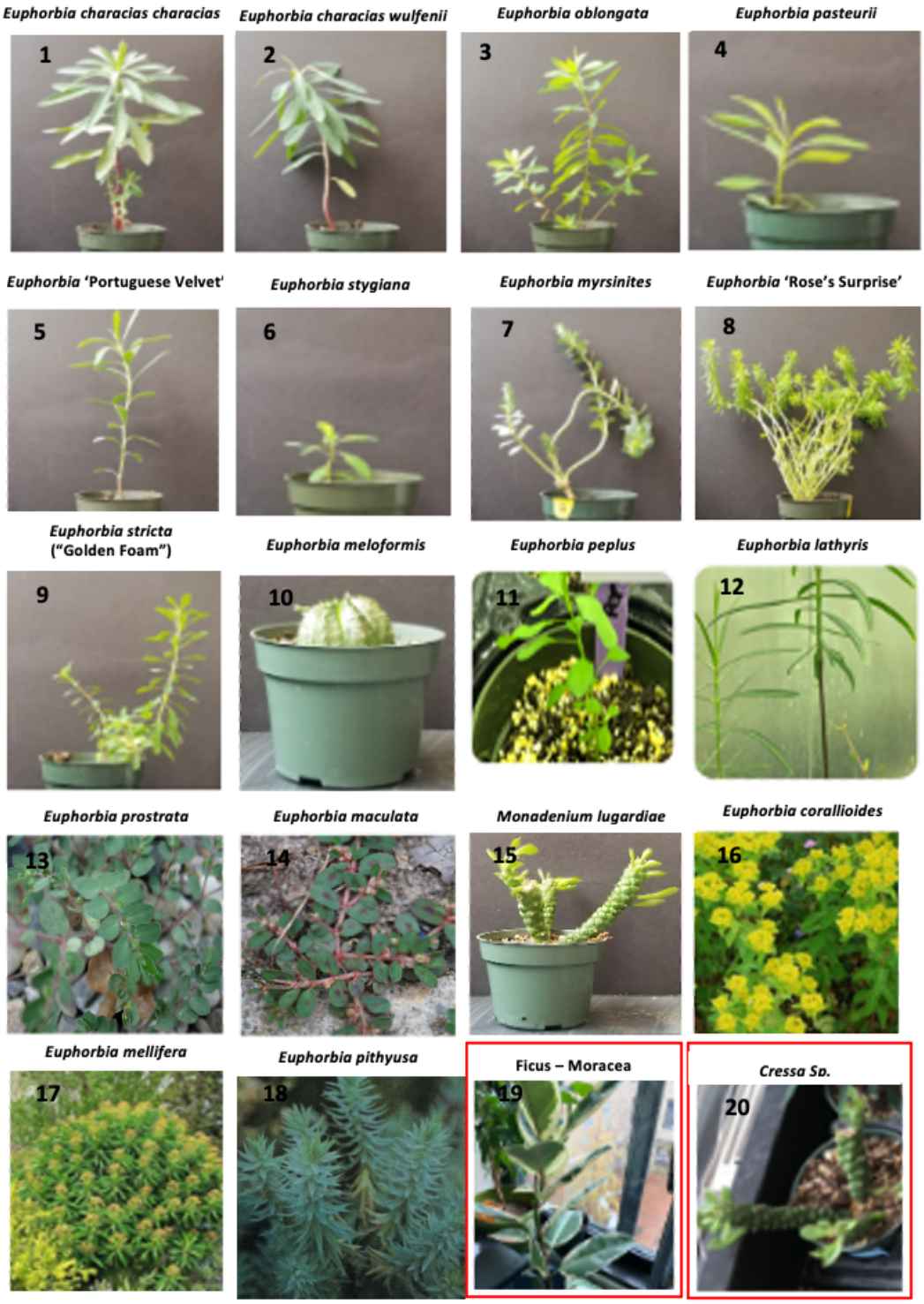

Figure 2. Representative pictures of each species studied.
18 *Euphorbia* species and two outgroup species (in red boxes) were included. Plant 14 is from the plantsam Plant Identification website (Frau-Doktor).

Extensive filtering of the MS1 LC-MS data was performed (**Section 2.2**). Based on this analysis, a total of 133 secondary metabolites were found in all plant samples in the negative ion mode. On the other hand, the positive ion mode had 815 secondary metabolites present in all species. Unfortunately, MS/MS analysis was not performed, hence, further identification of these metabolites is not possible. Considering only *E. peplus*, there were more unique metabolites (metabolites only in *E. peplus*) in the negative ion mode (22 peaks) than the positive ion mode (5 peaks). Overall, the total number of metabolites found in all plants was more abundant in positive ion mode, but unique metabolites from each species were more abundant in the negative ion mode. Therefore, it was important to analyze both the positive and negative ion modes to have a complete understanding of the *Euphorbia* latex secondary metabolite complexity.

Another analysis was performed to estimate the total number of secondary metabolites present in each plant sample **(Figure 3**) in addition to analyzing the total number of metabolites that were common to all species. These secondary metabolites were not necessarily unique to any species. In the positive ion mode, protonated or alkali ([M+Na]+ or [M+NH4]+) adduct analyte molecules were generally observed in the mass spectra while in the negative ion mode deprotonated adduct analyte molecules [M-H]- were observed (Steckel and Schlosser, 2019). It is thus possible that some of the peaks identified through the MS1 analysis are potential adducts or isomers.

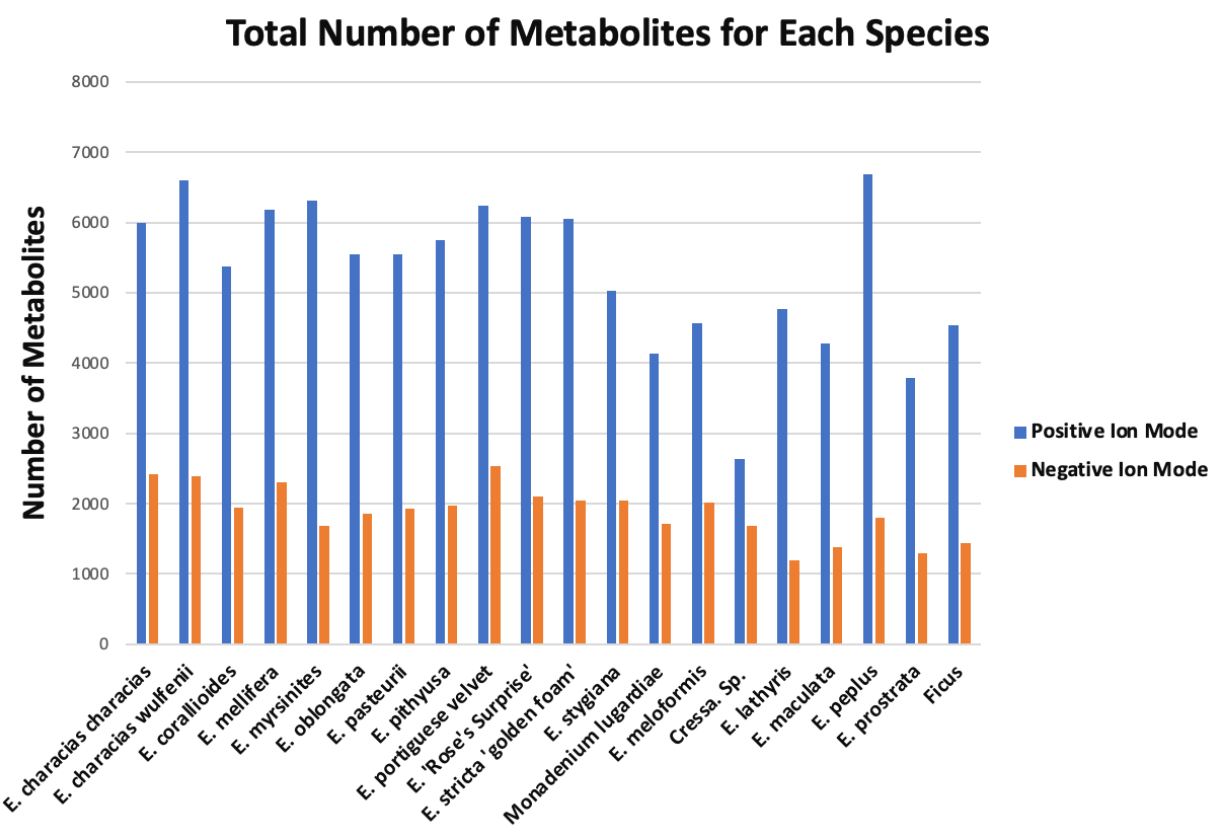

Figure 3. The total number of metabolites detected in each species
These bar graphs include 18 *Euphorbia* plant samples along with two outgroup plant samples. The numbers of metabolites with abundance greater than 10,000 was determined for each species. This process was repeated for all plant samples.

Lastly, the number of secondary metabolite peaks that are unique to each plant sample was also analyzed (**Figure 4**), since these peaks can inform us about latex secondary metabolite complexity in each plant in relation to other samples. Unlike the total number of secondary metabolites, more metabolite peaks were found in the negative ion mode. For both the positive and negative ion modes, *Euphorbia* Rose's Surprise had the most unique metabolites. There were some samples (*E. mellifera, E. prostrata,* etc.) without any unique metabolites either in one of the ion modes or both ion modes. Looking at the two outgroup plant samples, *Cressa* does not have unique metabolites, but *Ficus* had some unique metabolites. Although *E. peplus* did not have the most

abundant number of unique secondary metabolites, it has the greatest total number of metabolites in positive mode (**Figure 3**) and several unique metabolites (**Figure 4**). **Figure 5** shows a positive relationship between the total metabolite abundance and the number of unique metabolites in both the positive and negative ion modes. Therefore, after analyzing individual plants, 20 samples were compared amongst each other to better understand the *Euphorbia* latex complexity.

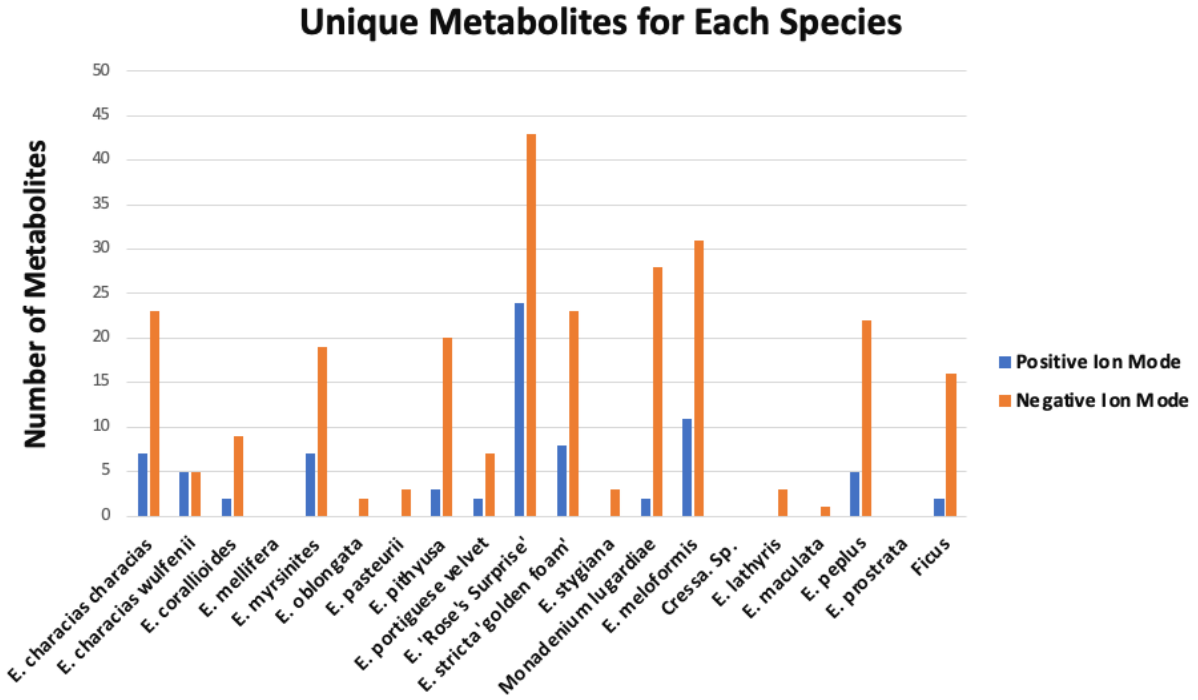

Figure 4. Number of metabolites unique to each species
These bar graphs include 18 *Euphorbia* plant samples along with two outgroup plant samples. Metabolites considered unique were present in one sample (with abundance greater than 10,000) and absent in all other samples (abundance less than 1,000).

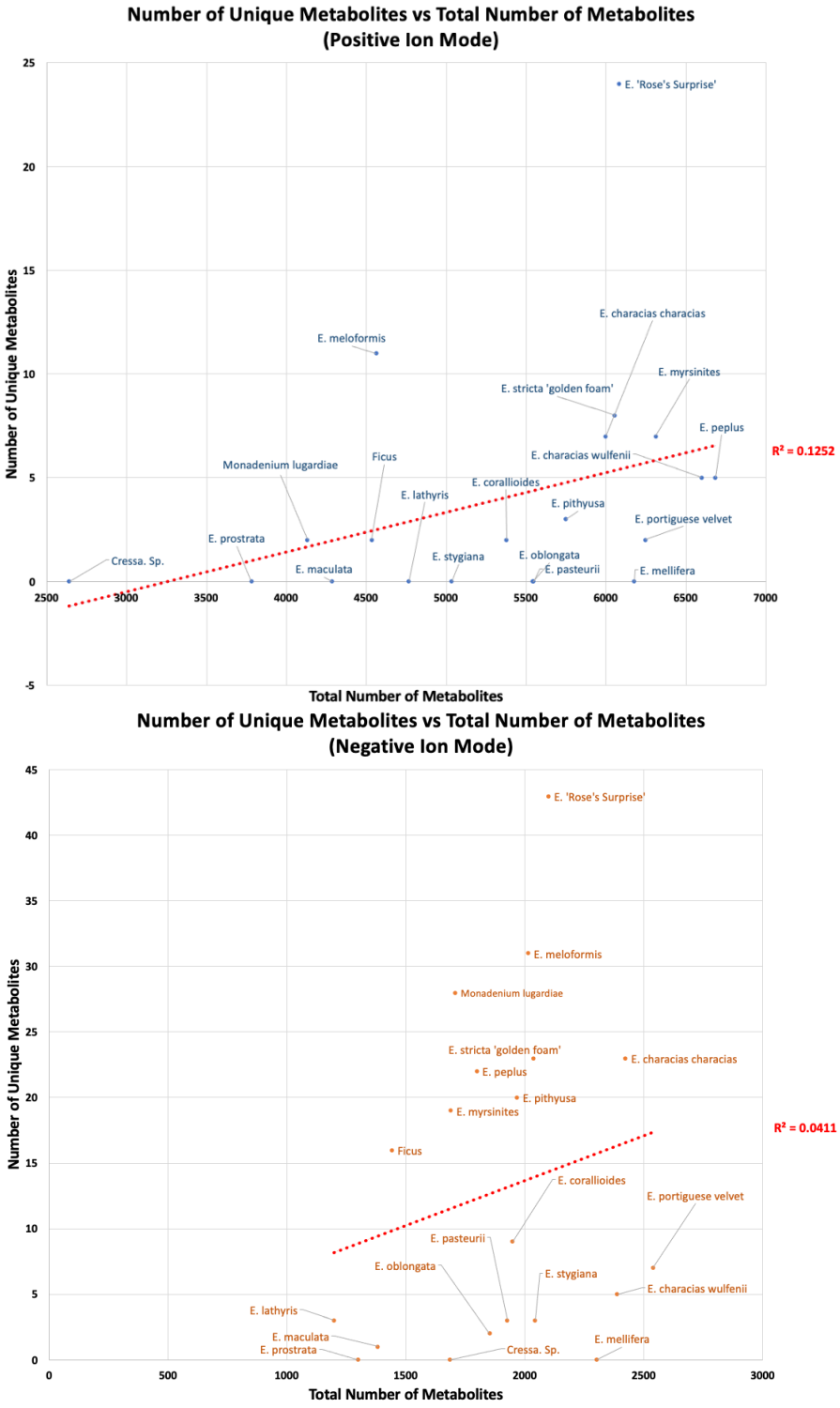

Figure 5. Relationship between the total number of metabolites and number of unique metabolites in both ion positive and negative ion modes

These scatterplots have positive R^2 values and indicating that there is a positive relationship between the total number of metabolites and the number of unique metabolites.

In addition to identifying unique secondary metabolites for each plant sample, principal component analysis (PCA) was performed for all 20 plant samples both in positive and negative ion modes (**Figure 6**) to identify metabolically similar groups of species. The PCA analysis clustered these plant species based on metabolomic abundance. For both ion modes, *E. peplus* was clustered together with *E. lathyris*, *E. mellifera, E. stricta* 'golden foam', and *Ficus*. However, *E. peplus* was relatively distant from other *Euphorbia* plant species, suggesting a more distinct latex profile. These species are evolutionarily distant from each other (Horn et al., 2012), lying in separate sub-clades of the genus, suggesting significant plasticity in latex profiles across the family. Since the *Euphorbia* genus has ~2,000 species scattered throughout four different subgenera including *Esula* (houses *E. peplus* and *E. lathyris*), *Rhizanthium*, *Euphorbia*, and *Chamaesyce*, it is possible that secondary metabolite diversity also varies substantially among species in the same subgenus (Horn et al., 2012). Shannon Entropy analysis was further performed to further visualize the complexity and secondary metabolite abundance of each plant sample.

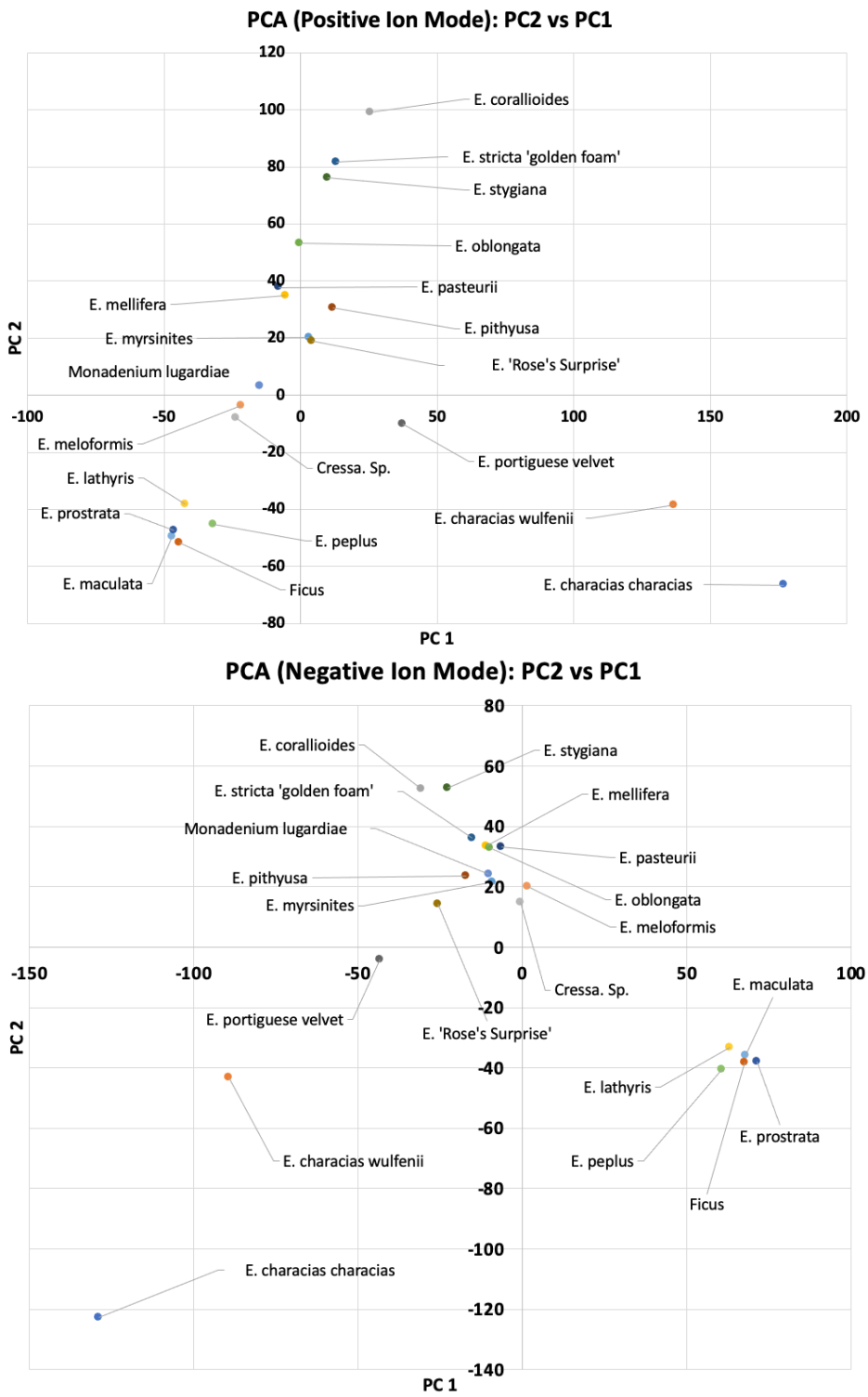

Figure 6. Principal component analysis in positive and negative ion modes
These graphs include 18 *Euphorbia* plant samples along with two outgroup plant samples. The PCA clustered these species based on metabolite abundancies.

The Shannon Entropy analysis was also performed in the positive and negative ion modes based on metabolite abundance (**Figure 7**) to quantify the complexity and diversity of the MS1 profile signal. Previously, Shannon Entropy was defined as a probability distribution over a finite number of samples (Shu-Cherng and Tsao, 2001). Currently, Shannon Entropy is used to measure the uncertainty related to a random variable (Wu et al., 2013). For this analysis, the big Shannon Entropy value indicates many different peaks that have different m/z values.

The *Ficus* and *Cressa* outgroup plant samples had fewer peaks with high entropy, which means that these two outgroup species' secondary metabolites have low abundance (number of peaks), while having many different peaks with different m/z values (Shannon Entropy). *E. characias characias* was the most diverse plant sample as there were many peaks with different m/z values, abundancies, and retention times. On the other hand, *E. prostrata* and *E. maculata* were the least diverse and least abundant sample, as these samples had relatively fewer peaks and lower Shannon Entropy values, which indicates that these fewer number of peaks are similar to each other in terms of the retention time, abundance, or m/z. However, based on these particular Shannon Entropy analysis results, species with high Shannon Entropy with most peaks cannot be said to have the most unique secondary metabolites as this Shannon Entropy only measures MS1 signals. In order to measure the uniqueness of the metabolites, MS2 data would have to be collected. Based on the MS1 Shannon Entropy result, *E. peplus* secondary metabolites were more diverse in the positive ion mode, while *E. lathyris was more diverse* in negative mode.

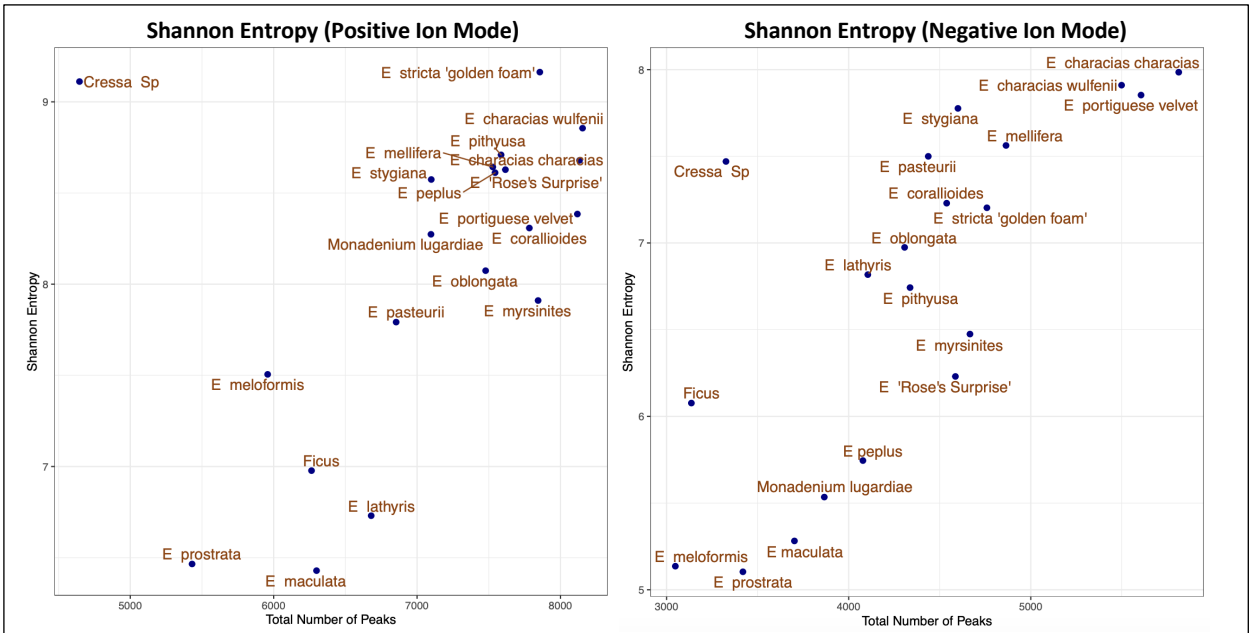

Figure 7. Shannon entropy analysis in positive and negative ion modes
These graphs include 18 *Euphorbia* plant samples along with two outgroup plant samples. Shannon entropy measured the diversity of metabolite abundancies in each species' metabolome. The x-axis charts the numbers of peaks with non-0 abundance per sample.

3.3 Identifying a model species for the *Euphorbia* genus, and estimating the genome sizes of *E. peplus and E. lathyris*

Despite being one of the largest angiosperm genera, a model species is not identified in the *Euphorbia* genus. In order to establish a model species for this project, two *Euphorbia* species were studied. *E. peplus* and *E. lathyris* were selected based on their relatively short life cycle compared to other potential *Euphorbia* model species. In addition, these two *Euphorbia* species are successful with their germination and seed production (Scientific Data Curation Team, 2020). *E. peplus* is found worldwide due to its high invasive potential (Hua et al., 2017). Similarly, *E. lathyris* is also found worldwide (Yu et al., 2020). *E. peplus* has been studied in the biochemical analysis, such as analysis of *E. peplus* latex lipase (Lazreg Aref et al., 2014). *E. lathyris* has been studied for its seed oil diterpenoids (Luo et al., 2016) and laticifers and latex (Castelblanque et al.,

2020). Given several characteristics described above and available information from previous studies, it makes *E. peplus* and *E. lathyris* suitable as a potential model species.

These selected species were first assessed to confirm their identities. Using sequence fragment from the *maturase K* (*matK*) gene. The *matK* genes found in the chloroplasts serve as a barcode for land plants for species identification (Yu et al., 2011). The *matK* sequence extracted from *E. peplus* and *E. lathyris* in the Moghe lab seed stock had 96% to 100% identity when compared with the published NCBI sequences (**Table 6**). Also considering their overall morphological features when compared to their descriptions, we verified that the plant samples were correct species. As mentioned by the Global Biodiversity Information Facility (GBIF), *E. peplus* and *E. lathyris* plants from the lab were herbaceous (Pava et al., 2017; Occdownload Gbif.Org, 2020). Furthermore, *E. peplus* and *E. lathyris* from the lab also have simple leaves, which is consistent with the description from the Encyclopedia of Life (Parr et al., 2014). These results confirmed the identity of the two species being investigated, allowing further analysis of the two species.

Table 6. *E. peplus* and *E. lathyris* NCBI *matK* sequences

Plant Name	***matK* Sequence**
E. peplus	AATACCTTACCCCATCCATCTCGAAAAATTGGTTCAAACCCTTC GTTATTGGGTGAAAGACCCTTCTTCTTTACATTTTTTACGATTCT TTCTTCATCAGTATTCGAATTTGAGCAGTCTTATTATTCCAAAG AAATCAATTTATTTTTTTCGAAAAAGGAATCCAAGGTTTTTCTT ATTCCTATATAATTCTCATATAACTGAATATGAATCCATCTTAT TTTTTCTCCGTAATCAGTCCTTTCATTTACGATCAACATTTTCTC GAGTCTTTCTTGAACGAATTTTTTTCTATGGAAAAATAGAACAT TGTGCAGAAGTTTTTGCTAATGATTTTCAGACCATTCTATTGTT GTTCAAGGATCCTTTGATGCATTATGTTAGATATCACGGAAAAT CCATTCTCGCTTTAAAAGATAAACCCTTTCTGATGAAAAAATG GCAATATTACCTTATTAATTTATCTCAATGTCATTTTTATGTCTG GTTTCAACCAAAAAAGATCTATATAAATTCATTATCAAAAAAT TCTCTCAATTTTTTGGGCTATCTTTCAAGTGTACAAAAGAATCC TTTGGTAGTACGGAGTCAAATGCTAGAAAATTCATATCTCATA

	GATAAAGAGAATACTATGAAGAAACTAGATACAATAGTTCCAA TTATTCCTTTAATTGGATTATTGTCAAAAATGAAATTTTGTAAC GCAGTGGGACATCCTATTAGTAAACCGATTCGGGCTCATTTATC CGATTCTGATATTATCGACCAATTTGCGCGTATATGTCGAAA
E. lathyris	TATGTGTCAGATGTATTAATACCGTACCCCATCCATCTCGAAAA ATTGGTTCAAACCCTTCGTTATTGGTTGAAAGATCCTTCTTTTTT GCATTTTTTACGACTCTTTCTTCATCAGTATTGGAATTGGAGCA GTCTTATTATTATAAATAAATCCATTTCTTTTTTTCGAAAAAGT AATCCAAGATTTTTCTTGTTCCTATATAATTCTCATATAGATGA ATATGAAGCCATCTTCTTTTTTCTCCGTAATCAGTCCTTTCATTT ACGATCAACATTTTCTCGAGTCTTTCTTGAACGAATTTTTTTCTA TGGAAAAATAGAACATTTTGCAGAAGTTTTTACTAATGATTTTC AGACCATTTTATGGTTGTTCAAGGATCCTTTCATGCATTATGTT AGATATCAAGGAAAATCAATTCTGGCTGTAAAAGATAAGCCCT TTCTGATGAAAAAATGGAAATATTATCTTGTCAATTTATGTCAA TGTCATTTTTATGTCTGGTTTCAACCAGAAAAGAGCTATATAAA TTCATTATCAAAAAATTCTCTCAACTTTTTGGGCTATCTTTCAA GTGTACAAAAAAATCCTTTGGTAGTACGGAGTCAAATGCTAGA AAATTCATATCTAATAGATAAGGATAATACTATGAAGAAACTC GATACAATAGTTCCAATTATTCCTTTAATTGGATTATTGGCAAA AACGAAATTTTGTAACGTAGTGGGACATCCTATTAGTAAACCG ATCCGGGCTCATTCATCAGATTCTGATATTATCGACCAATTTGC GCGTATATGTAGAAATTTTGCTCATTTTT

After verifying *E. peplus* and *E. lathyris*, flow cytometry analysis was used to estimate the genome size, which was not yet studied to our knowledge. The flow cytometry analysis (**Figure 8**) indicates that *E. peplus* had a smaller average 2C genome size of 1.13 pg while *E. lathyris* had a larger average 2C genome size of 4.50 pg (**Supplement Part 4**). Assuming 1 pg corresponds to 1,000 Mb (Gregory, 2005), we estimated the diploid 2C genome sizes of the two species as 1,130 Mb and 4,500 Mb, respectively. These values were later scaled to obtain the correct genome sizes (see Discussion).

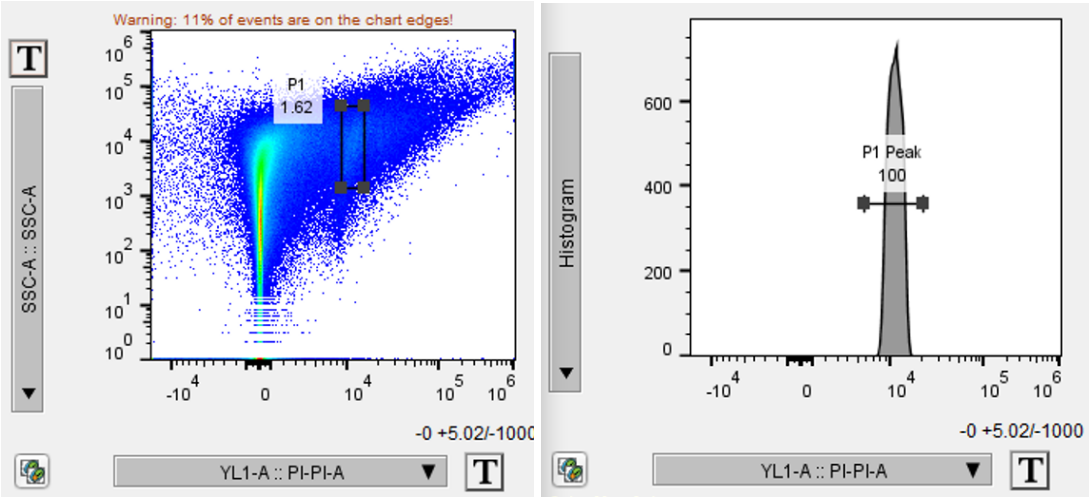

Figure 8. Flow cytometry analysis graphs for *E. peplus* Trial 1 (**Supplement Part 4**) P1 Peak from the right graph indicates the histogram formed by the gate surrounding P1 region from the left graph. The peak area from this histogram was used to calculate genome sizes.

E. peplus was thus selected as a model species for genome sequencing due to the following reasons: (i) compared to *E. lathyris*, *E. peplus* seeds were easier to germinate and the plants grew faster, (ii) *E. peplus* took less time to mature and produce seeds compared to *E. lathyris,* (iii), a smaller genome size makes it easier to sequence and transform *E. peplus*. Other members of the Moghe and Frank labs continued this experiment by extracting RNA and DNA from the two species, with only *E. peplus* DNA submitted for genome sequencing and RNA from both species submitted for RNA-seq. To continue a more in-depth analysis of the metabolomes of the two species, extensive MS/MS characterization of their latex was performed.

3.4 *E. peplus* and *lathyris* latex specialized metabolite diversity

Latex contains diverse classes of secondary metabolites, which often serve as a plant defense mechanism. *E. peplus* and *E. lathyris* latex were extracted using the optimized protocol described in Section 2.2 and analyzed using the MS1 and MS2 results. Extracted latex was analyzed using Canopus and Sirius to obtain structural annotations, and the numbers of metabolite

"Superclasses" were quantified to determine the different types of metabolites present in the *E. peplus* and *E. lathyris* latex. After predictions for three compounds for each mgf entry were made by Sirius, Canopus assigns the class to each of those three compounds. The *Euphorbia* latex from *E. peplus* and *E. lathyris* has a relatively high amount of benzenoids, lipids and lipid-like molecules, and organic nitrogen compounds (**Figure 9**). The lipids and lipid-like molecule superclass houses terpenes and terpene derivatives. These secondary metabolites were analyzed to determine the number of metabolites that are significantly more abundant in each species.

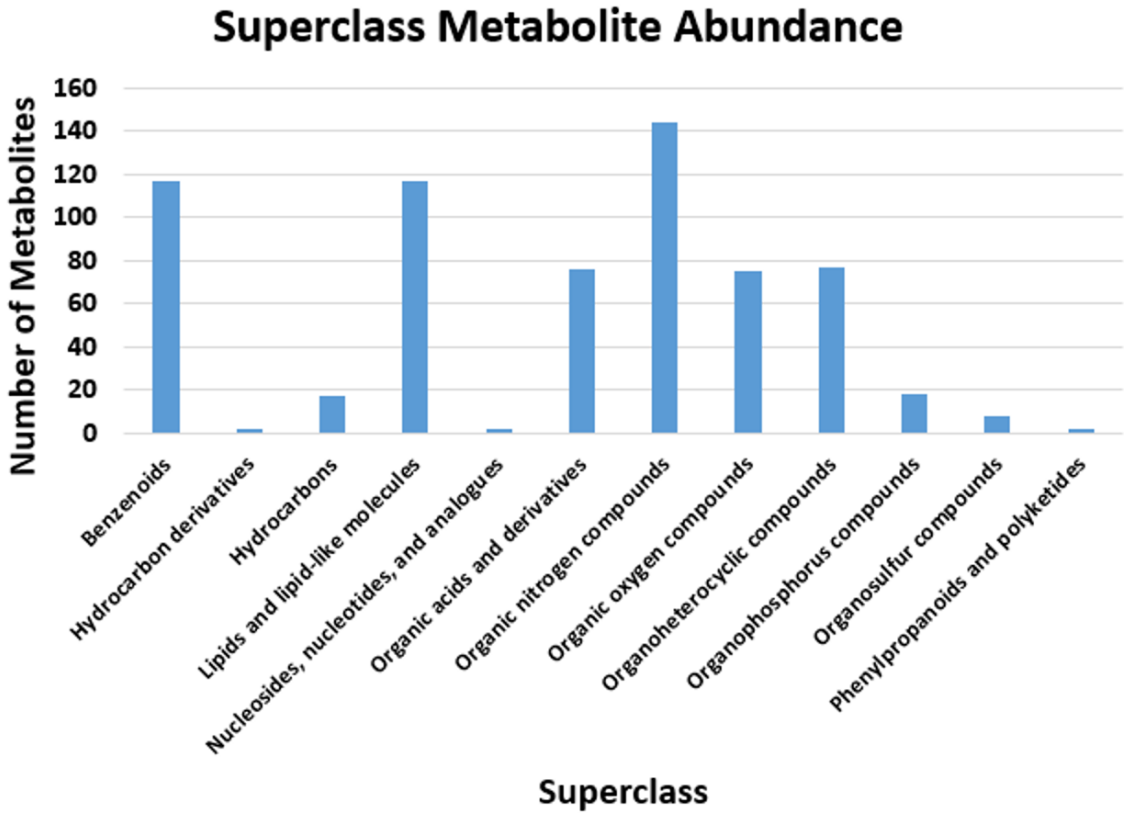

Figure 9 Numbers of metabolites assigned to different superclasses based on annotation. Each detected metabolite was annotated by its superclass based on Canopus and Sirius. Superclass represents specific group of metabolites based on their chemical structures.

All secondary metabolites from these superclasses were visualized using a volcano plot in order to identify metabolites that are more abundant in either *E. peplus* or *E. lathyris* based on metabolites that are common to both species. A total of 873 secondary metabolites were more than two-fold abundant (with a corrected $p<0.05$) in either *E. peplus* or *E. lathyris* and 2785 metabolites that were evenly found in both species (**Figure 10**). Such a result indicates that a relatively high number of metabolites are conserved through different *Euphorbia* species. In addition, *E. peplus* has 685 secondary metabolites while *E. lathyris* only has 187 secondary metabolites that had a corrected p-value of less than or equal to 0.05, which was used for the t-test analysis. After quantifying all metabolites, unique terpenes from these two species were analyzed.

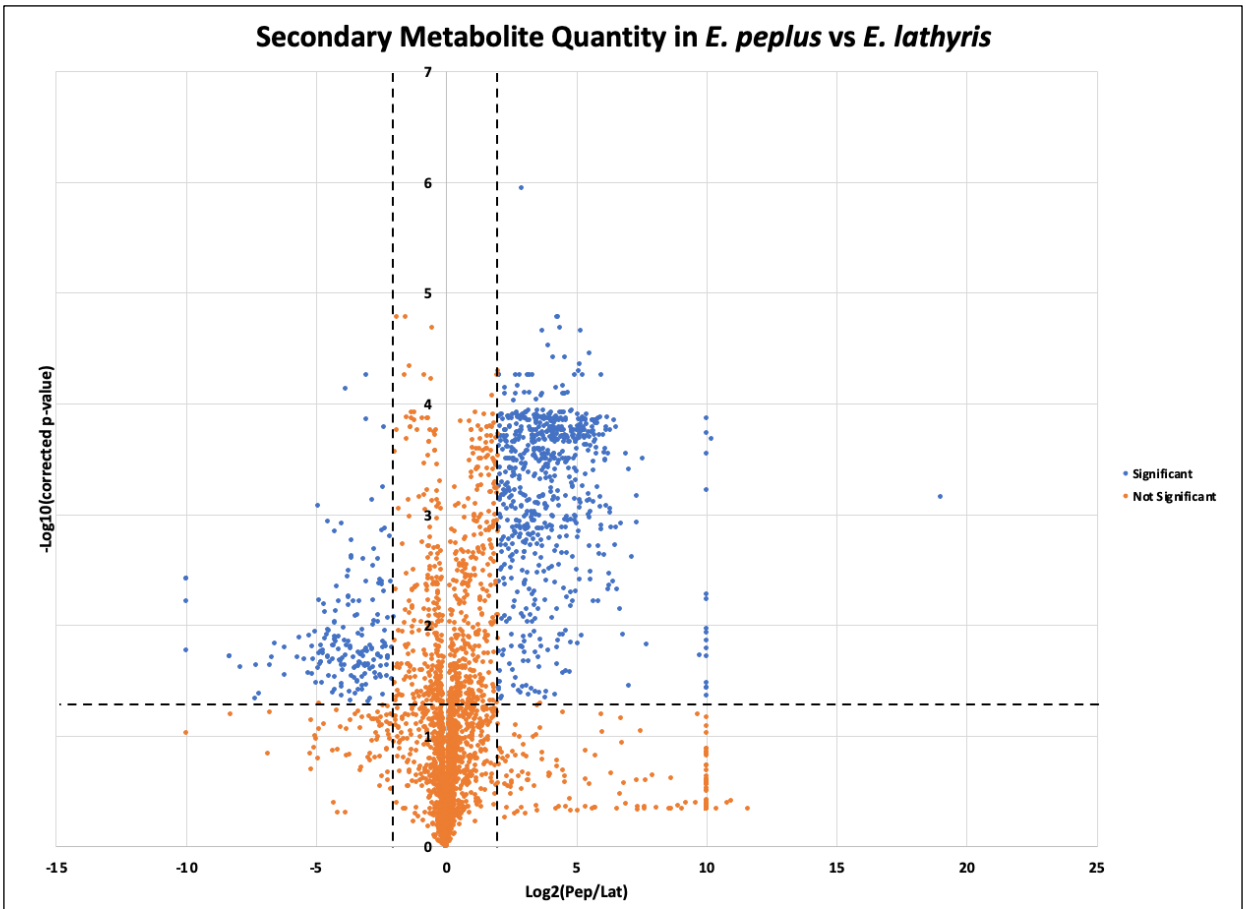

Figure 10. Secondary metabolite quantity in *E. peplus* vs *E. lathyris*
Six *E. peplus* and three *E. lathyris* plants were used. The x-axis represents $\log_2$(average abundance in peplus/average abundance in lathyris) and the y-axis represents $-\log_{10}$(corrected p-value). The p-values were corrected with Benjamini Hochberg correction method. Pep is an acronym for *E. peplus,* and Lat is an acronym for *E. lathyris.* The corrected p-value was taken from the two tailed t-test with unequal variance. Any y-values > 1.30103 (= corrected p-value of 0.05) and any x-values > 2 or < -2 are considered to be significant secondary metabolites. The points with -10 or +10 x-values represent secondary metabolites that are only present in *E. peplus* or *E. lathyris* respectively.

Terpenes are one of the most important secondary metabolites from *Euphorbia* latex (Nemethy et al., 1983). Out of 16 metabolites identified as terpenes from *E. peplus* and *E. lathyris* latex, three terpenes were noticeably more abundant (corrected $p \leq 0.05$) in either species (**Figure 11**). Only one of these terpenes was present in a greater amount for *E. peplus,* and two terpenes

were present in a greater amount for *E. lathyris*. Other than these, 13 terpenes were similar in quantity for both plants. However, a lower total number of terpenes was identified than what was expected. This terpene abundance is based on the Sirius and Canopus identification and there could be more terpenes with other functional groups that were not identified by these programs as terpenes. Another metabolome analysis approach called molecular networking was used to connect LC-MS/MS peaks based on the cosine score similarities of the fragmentation patterns. This strategy can also be used to putatively identify metabolites whose identification cannot be readily obtained.

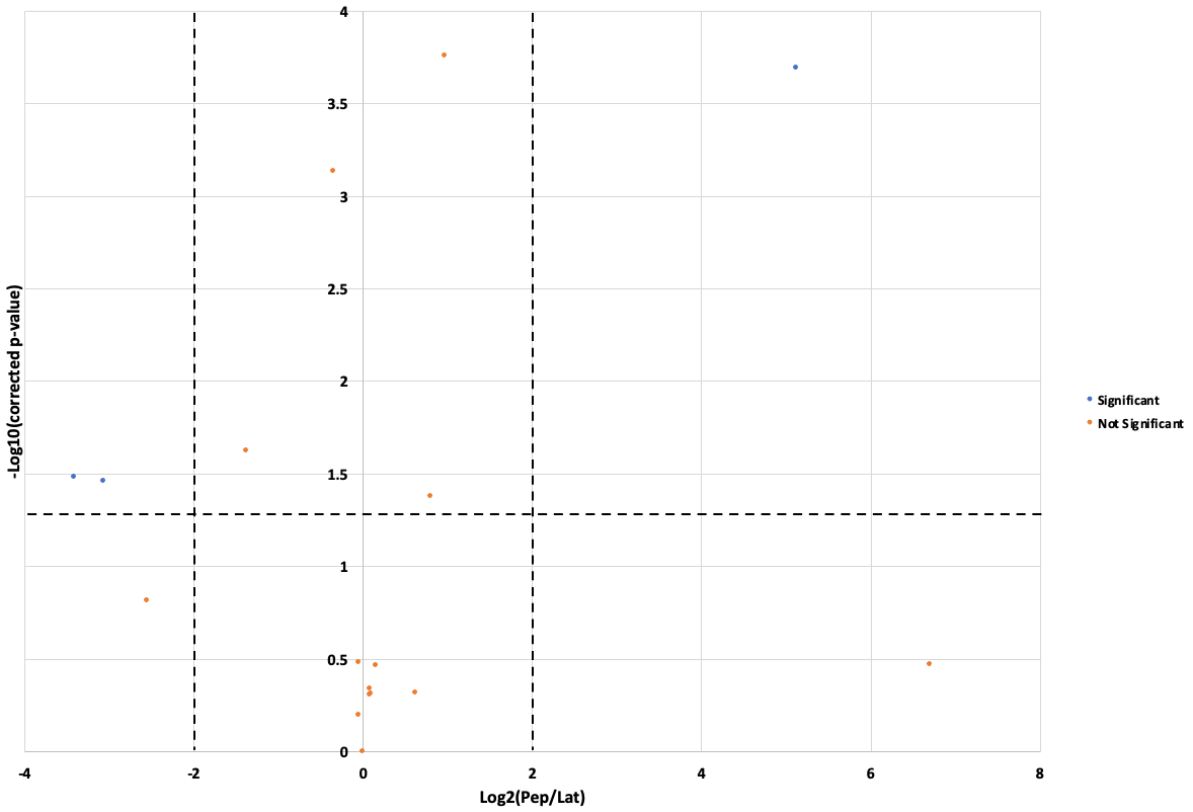

Figure 11. Abundancies of terpenes in *E. peplus* vs *E. lathyris*
Six *E. peplus* and three *E. lathyris* were plants used. The x-axis represents $\log_2$(average abundance in peplus/average abundance in lathyris) and the y-axis represents $-\log_{10}$(corrected p-value). The p-values were corrected with Benjamini Hochberg correction method. Pep is an acronym for *E. peplus,* and Lat is an acronym for *E. lathyris.* The corrected p-value was taken from the two tailed t-test with unequal variance. Any y-values > 1.30103 (= corrected p-value of 0.05) and any x-values > 2 or < -2 are considered to be significant terpenes.

Canopus used the MSP files to identify terpenes and prenol lipids and then Cytoscape was used to visualize the node and modified edge in the form of a molecular network using the Prefuse Force Directed Layout. Secondary metabolites that are identified as terpenes have nodes colored in pink (**Figure 12**). Despite their annotation similarity, the terpenes were not found clustered close to each other. There are three smaller networks (labeled 1, 2, and 3) with each containing a terpene. It is possible that the secondary metabolites that are directly connected with the pink nodes are

terpenes. To understand the types of secondary metabolites that were connected to the pink terpene nodes and to understand their relationships, superclasses were differentiated for the second molecular networking analysis.

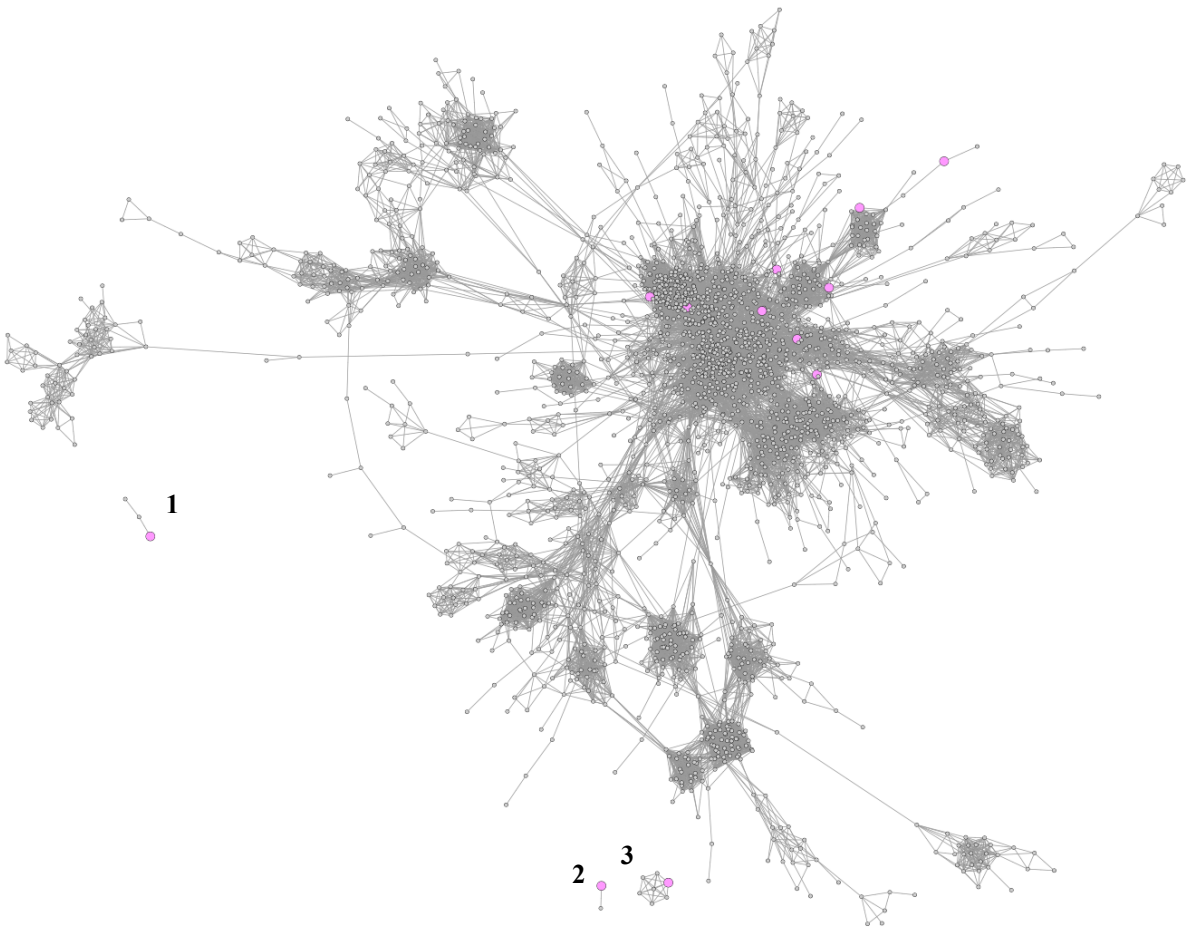

Figure 12. Molecular networking of terpenes
The nodes represent secondary metabolites, and the edges represent the cosine score of the metabolite fragment peaks. The nodes that are close to each other have similar fragmentation patterns. The nodes that are larger in sizes with pink color represent identified terpenes.

A second molecular network was created by applying blank and adduct filters in order to remove metabolites that were potentially in-source fragments and thus, include only non-redundant metabolites. Light green nodes with pink borders represent identified terpenes (**Figure**

13). Molecules that were identified as terpenes were scattered rather than being close to one another as shown by **Figure 12**, but more implications about the secondary metabolites that were connected to terpenes could be made based on their superclasses. Terpenes could be part of the alkaloid group with at least one nitrogen (i.e. diterpene aconitine) (Guggisberg and Hesse, 2003), or other structures collectively termed mero-terpenoids (Gozari et al., 2021). Therefore, the nodes that were labeled as organic nitrogen compounds could be terpene molecules if these metabolites are connected to the identified terpenes. If all terpenes were clustered together as a group, then it is most likely that there were only 12 identified terpenes. However, since all identified terpenes were not connected in this network, it would be possible that there could be more terpenes than what was labeled as terpenes. These potential terpenes could have functional groups that resemble other superclasses. To visualize these potential functional groups, superclasses were assigned for the second molecular networking analysis. TPS sequences were selected for further studies as these are specialized enzymes that produce terpenes, which is one of the major components of *Euphorbia* latex. TPS sequences from Euphorbiaceae plants along with *E. peplus* and *E. lathyris* in relation to other plants' TPS sequences were analyzed to understand the evolutionary and phylogenetic relationship of *Euphorbia* TPS.

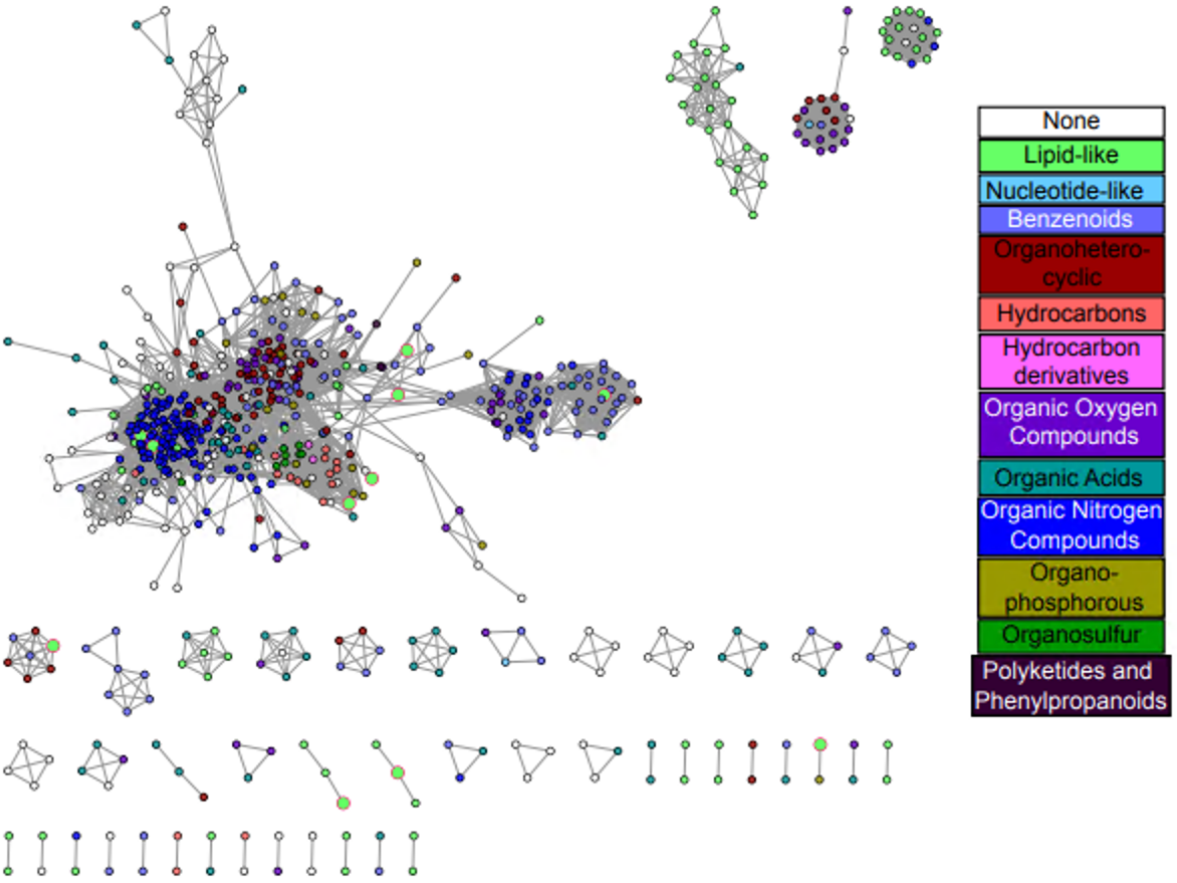

Figure 13. Molecular networking of terpenes (blank and adduct filtered)
The nodes represent secondary metabolites, and the edges represent the cosine score of the metabolite fragment peaks. The nodes that are close to each other have similar fragmentation patterns. The nodes are colored according to their Canopus-predicted Superclass. The nodes representing terpenes have pink borders and green color and are larger in sizes.

3.5 *E. peplus* and *lathyris* terpene synthases are abundant and conserved

Since TPS sequences are responsible for producing these identified unique terpenes (**Section 3.4**), TPS sequences were analyzed in order to understand the correlation between the terpene production and the number of terpene synthases. These TPS were identified based on the genome and transcriptome protein FASTA sequences and hmmsearch analysis. The number of TPS found for each Euphorbiaceae species is listed in **Table 7**. These TPS sequences were used

to develop a phylogenetic tree. This phylogenetic tree maintained all previously known TPS groups (a, b, c, e, f, g, and h) (Chen et al., 2011). Euphorbiaceae genome-based TPS sequences and *Euphorbia* transcriptome-based terpene synthase sequences were added to the phylogenetic tree (**Figure 14**). The TPS-d clade was not present on this phylogenetic tree because gymnosperm terpene synthase sequences were not included as this project is looking at angiosperms. Even with the same terpene synthase subfamilies, there were some organizational differences from the published paper (Chen et al., 2011). **Figure 14** presented a larger a-1 subgroup within the TPS-a subfamily and that the subfamilies TPS-e and TPS-h were grouped instead of TPS-e and TPS-f. The BLAST search was performed to understand the different organization of the phylogenetic tree (**Figure 14**) compared to the published tree. A custom BLAST search was used to verify published terpene synthase sequences (Chen et al., 2011) with the terpene synthase sequences generated by hmmsearch for *Arabidopsis*, *Oryza*, *Physcomitrella*, *Populus*, *Selaginella*, *Sorghum*, and *Vitis.* Most TPS-a sequences matched, but there were some differences for other TPS subfamilies. Some TPS sequences that were published as TPS-b were found in TPS-g and some TPS-a sequences were found in TPS-b. Overall, even with such diverse secondary metabolites, especially terpenes, *E. peplus* and *E. lathyris* had conserved terpene synthases in relation to other plants. There were no novel clades formed by TPS' in the *Euphorbia* species.

Table 7. Number of terpene synthases present in both N-terminus and C-terminus domains

Plant Name	Number of Terpene Synthases
Heavea brasiliensis (genome)	56
Jatropha curcas (genome)	61
Manihot esculenta (genome)	45
Ricinus communis (genome)	32
E. peplus (transcriptome)	89
E. lathyris (transcriptome)	46

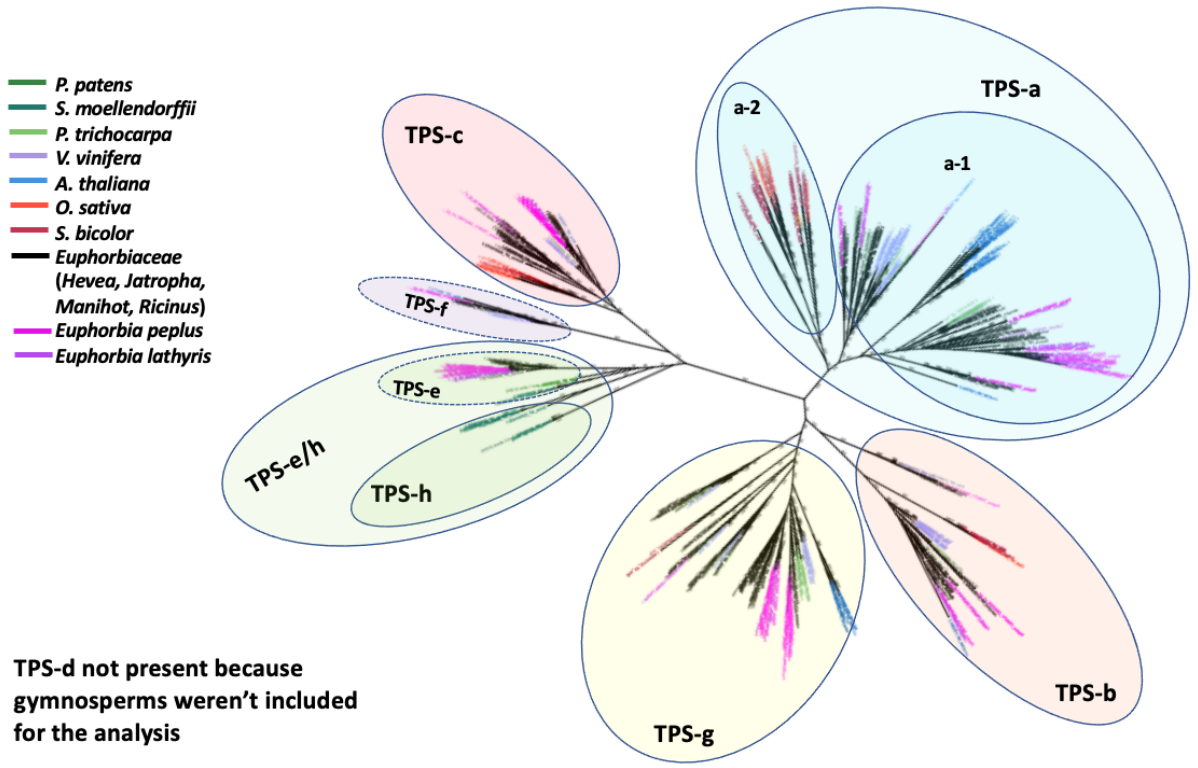

Figure 14. Phylogeny of terpene synthases (TPS)
The TPS subfamilies were grouped by comparing with the published phylogenetic tree (Chen et al., 2011).

4. DISCUSION

4.1 From a genome perspective, Euphorbia peplus is a good model system for further studies

E. peplus was finally selected as the model species based on its smaller genome size, compared to *E. lathyris* genome size, measured by the flow cytometry analysis. According to two published articles (Santana et al., 2016; Loureiro et al., 2007), the measured genome size of *E. peplus* is 1C = 0.35 pg. This published value is slightly less than the value calculated based on our flow cytometry of 1C = 0.565 pg, which corresponds to a genome size of ~552 Mb (given 1 pg = 978 Mb). However, the calculated average haploid 1C genome size of *Solanum lycopersicum* (1460 Mb) was higher than the published 1C value (~920 Mb) (Barone et al., 2008; Michaelson et

al., 1991). Multiplying our estimated *E. peplus* 1C genome size with a factor of 0.63 derived from this ratio (= [tomato published]/[tomato calculated]), we can re-estimate the *E. peplus* 1C value as 337 Mb. Preliminary results from PacBio sequencing provide a 1C genome size estimate of ~332 Mb. The sequencing-based estimate is thus supported by the flow cytometry estimate. The scaling factor, when applied to the *E. lathyris* flow cytometry results, gives an estimated genome size of 1386 Mb.

The main purpose of the flow cytometry analysis was to find the *Euphorbia* species with a smaller genome size. Since *E. lathyris* had an average genome size that was four times greater than that of *E. peplus*, as a model species, *E. peplus* was selected. After verifying *E. peplus* and *E. lathyris* plants based on the *matK* analysis and genome sizes were measured, these two species were used for quantifying the latex secondary metabolites.

4.2 A computational analysis of latex secondary metabolites and terpene diversity

Based on the MS2 metabolomics analysis of *E. peplus* and *E. lathyris* latex, relatively fewer secondary metabolites were identified as terpenes than what was expected. There were only 16 identified terpenes (labeled as prenol lipids and steroids and steroid derivatives) that survived filters and thresholds. Comparing with previous studies, one article states that terpenes, especially triterpenes, are the major components of the *Euphorbia* latex (Nemethy et al., 1983). Approximately half of the dry latex weight covers triterpenes and their derivatives. *E. lathyris* was used for the latex analysis. On the other hand, another article states that a total of 13 terpenes were profiled from the *E. peplus* latex (Hua et al., 2017), which was similar to the number of terpenes identified from our experiment. According to this paper, two new diterpenoids were found in addition to previously identified 11 terpenes. These known terpenes include ten diterpenes and

acyclic triterpene alcohol peplusol and these were analyzed using HPLC and UPLC-MS/MS. Even with only 13 terpenes, the result from this paper illustrates that *E. peplus* latex terpenes are diverse (Hua et al., 2017). Such varying results may be due to the sample size and experimental method. Since only *E. peplus* and *E. lathyris* latex were analyzed using MS/MS, it was difficult to conclude on the overall *Euphorbia* latex terpene diversity. For our result, the restrictions and filters that were applied could have been too strict, and that there could be more terpenes than that were identified, especially based on the molecular networking analysis (**Figures 12 and 13**). Based on **Figure 13**, there may be more terpenes than ones that are identified as some could contain other functional groups such as nitrogen. Further analysis could be done using different *Euphorbia* samples, not just limited to *E. peplus* and *E. lathyris* while increasing the samples for each species to better understand the terpene diversity.

4.3 Euphorbiaceae TPS diversity in relation to other previously studied species' TPS

TPS sequences are partly responsible for the observed terpenoids diversity in plants. The number of terpene synthases found in both the C-terminus and N-terminus *Ricinus communis* and *Manihot esculenta* was similar for the published and the hmmsearch results (Jiang et al., 2019). The published result had 23 and 41 TPS for *Ricinus communis* and *Manihot esculenta* respectively, while the hmmsearch result had 32 and 45 TPS in our study. This shows that proper parameters were set for the hmmsearch analysis. However, since genome sequences for *E. peplus* and *E. lathyris* are not available, the transcriptomes developed by the Moghe Lab were used to perform the hmmsearch analysis. The same parameters were used when performing the phylogenetic analysis with these two *Euphorbia* transcriptomes. *E. peplus* had a relatively large number of TPS (89) compared to other Euphorbiaceae species (**Table 7**). This makes it impossible to directly

compare terpene synthase diversity for the *Euphorbias*. However, using these transcriptome results, TPS subfamilies were analyzed in relation to *Hevea brasiliensis, Jatropha curcas, Manihot esculenta*, and *Ricinus communis* along with other published species (Chen et al., 2011). Overall, the reconstructed phylogenetic tree had a similar structure to the published result. However, there were some discrepancies between subfamilies assigned by the published paper and phylogeny tree that was developed as terpene synthase published as TPS-b were found in TPS-g and some published TPS-a sequences were found in the TPS-b subfamily. There are several reasons for such discrepancies. The subfamilies were divided and grouped based on branching and the color code. TPS-a, TPS-b, and TPS-g are closely located, so some sequences might have been grouped with different TPS subfamilies that are nearby. Furthermore, terpene synthase sequences were obtained using different methods, so the sequences used might not have been entirely the same. For the future experiment, it would be helpful to use the same terpene synthase sequences from the published result and prepare a phylogenetic tree with Euphorbiaceae and *Euphorbia* terpene synthases. In addition, when genomes are available for *E. peplus* and *E. lathyris,* terpene synthase sequences should be analyzed again for a more accurate analysis.

Overall, based on the *Euphorbia* transcriptome, the *Euphorbia* TPS sequences seem to be conserved and these diverse TPS subfamilies could have been related to diverse terpenes produced by the *Euphorbia* latex. However, we found that TPS subfamilies TPS-h and TPS-a-2 were missing in the *E. peplus* transcriptome and the *E. lathyris* transcriptome. These subfamilies have been previously described to have sesquiterpene synthase function (TPS-a-2) and putative bifunctional diterpene synthase function (TPS-h) respectively (Chen et al., 2011). Since there are known diterpenes for *E. peplus* and *E. lathyris*, it would be possible that these subfamilies were missing due to an incomplete transcriptome.

5. CONCLUSION

In this study, we isolated over 3000 metabolite peaks from LC-MS data across 18 *Euphorbia* species. Machine learning analysis helped predict 16 terpenes from *E. peplus* and *E. lathyris* and revealed other metabolite classes such as benzenoids, hydrocarbons, lipid, and lipid-like molecules. These results illustrate how diverse the *Euphorbia* latex secondary metabolites are. Further studies would have to be conducted with various *Euphorbia* species to better understand latex metabolite structural diversity. This study also identified *E. peplus* as the model species for further genome analysis in the *Euphorbia* genus, with a haploid genome size of 337 Mb.

In addition, we found that the Euphorbiaceae and *Euphorbia* genus terpene synthases are present in all subfamilies except for TPS-h. Although *E. peplus* and *E. lathyris* terpene syntheses were based on transcriptomes, which made it difficult to assess the *Euphorbia* TPS diversity, we were able to see that the *Euphorbia* terpene synthases were present throughout different subfamilies. Further research performed using the foundation developed in this thesis will help illuminate whether a high level of latex metabolite diversity and numerous terpene synthases potentially allowed the *Euphorbia* genus to be one of the most species-rich genera in the plant kingdom.

ACKNOWLEDGEMENTS

I would like to recognize my PI Dr. Gaurav Moghe, Alexandra Bennett, Elizabeth Mahood, and Dr. Lars Kruse for supporting and encouraging me throughout my journey. Their advice was very helpful in moving through different phases of the project. While working on the Euphorbiaceae project and honors thesis, I learned a lot about scientific research and various research skills. Without their help, I would not have been made it this far. I also appreciate my friends and family for providing constant support. This research was funded by the Institute of Biotechnology Seed Grant given to Drs. Gaurav Moghe and Margaret Frank.

REFERENCES

Adusumilli, R. and Mallick, P. (2017). Data Conversion with ProteoWizard msConvert. In Proteomics, L. Comai, J.E. Katz, and P. Mallick, eds, Methods in Molecular Biology. (Springer New York: New York, NY), pp. 339–368.
Armbruster, W.Scott., Gillespie, L.J., and Smithsonian Institution. (1997). A contribution to the Guianan flora : Dalechampia, Haematostemon, Omphalea, Pera, Plukenetia, and Tragia (Euphorbiaceae) with notes on subfamily Acalyphoideae / (Smithsonian Institution Press,: Washington, D.C. :).
Arumuganathan, K. and Earle, E.D. (1991a). Estimation of nuclear DNA content of plants by flow cytometry. Plant Mol Biol Rep **9**: 415–415.
Arumuganathan, K. and Earle, E.D. (1991b). Nuclear DNA content of some important plant species. Plant Mol Biol Rep **9**: 415–415.
Aubourg, S., Lecharny, A., and Bohlmann, J. (2002). Genomic analysis of the terpenoid synthase (AtTPS) gene family of Arabidopsis thaliana. Mol Gen Genomics **267**: 730–745.
Barone, A., Chiusano, M.L., Ercolano, M.R., Giuliano, G., Grandillo, S., and Frusciante, L. (2008). Structural and Functional Genomics of Tomato. International Journal of Plant Genomics **2008**: 1–12.
Basak, S., Bakshi, P., Basu, S., and Basak, S. (2009). Keratouveitis caused by *Euphorbia* plant sap. Indian J Ophthalmol **57**: 311.
Boodley, J.W. and Sheldrake, R. (1972). Cornell Peat-Lite Mixes for Commercial Plant Growing. A Cornell Cooperative Extension Publication.
Bruyns, P.V., Mapaya, R.J., and Hedderson, T.J. (2006). A new subgeneric classification for *Euphorbia* (Euphorbiaceae) in southern Africa based on ITS and *psbA-trnH* sequence data. Taxon **55**: 397–420.
Castelblanque, L., García-Andrade, J., Martínez-Arias, C., Rodríguez, J.J., Escaray, F.J., Aguilar-Fenollosa, E., Jaques, J.A., and Vera, P. (2020). Opposing Roles of Plant Laticifer Cells in the Resistance to Insect Herbivores and Fungal Pathogens. Plant Communications: 100112.
Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom: Terpene synthase family. The Plant Journal **66**: 212–229.
Christenhusz, M.J.M. and Byng, J.W. (2016). The number of known plants species in the world and its annual increase. Phytotaxa **261**: 201.
Crepet, W.L. (1984). Advanced (Constant) Insect Pollination Mechanisms: Pattern of Evolution and Implications Vis-a-Vis Angiosperm Diversity. Annals of the Missouri Botanical Garden **71**: 607.
Demir, Y., Alayli, A., Yildirim, S., and Demir, N. (2005). Identification of Protease from *Euphorbia amygdaloides* Latex and its Use in Cheese Production. Preparative Biochemistry and Biotechnology **35**: 291–299.
Dudareva, N., Martin, D., Kish, C.M., Kolosova, N., Gorenstein, N., Fäldt, J., Miller, B., and Bohlmann, J. (2003). (*E*)-β-Ocimene and Myrcene Synthase Genes of Floral Scent Biosynthesis in Snapdragon: Function and Expression of Three Terpene Synthase Genes of a New Terpene Synthase Subfamily. Plant Cell **15**: 1227–1241.

Dufour, D.L. (1988). Cyanide content of cassava (Manihot esculenta, Euphorbiaceae) cultivars used by Tukanoan Indians in Northwest Amazonia. Econ Bot **42**: 255–266.
Ehrlich, P.R. and Raven, P.H. (1964). Butterflies and Plants: A Study in Coevolution. Evolution **18**: 586.
Ernst, M. et al. (2019). Assessing Specialized Metabolite Diversity in the Cosmopolitan Plant Genus Euphorbia L. Front. Plant Sci. **10**: 846.
Ernst, M., Grace, O.M., Saslis-Lagoudakis, C.H., Nilsson, N., Simonsen, H.T., and Rønsted, N. (2015). Global medicinal uses of Euphorbia L. (Euphorbiaceae). Journal of Ethnopharmacology **176**: 90–101.
Ernst, M., Saslis-Lagoudakis, C.H., Grace, O.M., Nilsson, N., Simonsen, H.T., Horn, J.W., and Rønsted, N. (2016). Data from: Evolutionary prediction of medicinal properties in the genus Euphorbia L.: 14295299 bytes.
Farrell, B.D., Dussourd, D.E., and Mitter, C. (1991). Escalation of Plant Defense: Do Latex and Resin Canals Spur Plant Diversification? The American Naturalist **138**: 881–900.
Foisy, M.R., Albert, L.P., Hughes, D.W.W., and Weber, M.G. (2019). Do latex and resin canals spur plant diversification? Re-examining a classic example of escape and radiate coevolution. J Ecol **107**: 1606–1619.
Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research **40**: D1178–D1186.
Gozari, M., Alborz, M., El-Seedi, H.R., and Jassbi, A.R. (2021). Chemistry, biosynthesis and biological activity of terpenoids and meroterpenoids in bacteria and fungi isolated from different marine habitats. European Journal of Medicinal Chemistry **210**: 112957.
Gregory, T.R. (2005). Genome Size Evolution in Animals. In The Evolution of the Genome (Elsevier), pp. 3–87.
Guggisberg, A. and Hesse, M. (2003). Alkaloids. In Encyclopedia of Physical Science and Technology (Elsevier), pp. 477–493.
Hagel, J., Yeung, E., and Facchini, P. (2008). Got milk? The secret life of laticifers. Trends in Plant Science **13**: 631–639.
Hillwig, M.L., Xu, M., Toyomasu, T., Tiernan, M.S., Wei, G., Cui, G., Huang, L., and Peters, R.J. (2011). Domain loss has independently occurred multiple times in plant terpene synthase evolution: Complex evolutionary origins for terpene synthases. The Plant Journal **68**: 1051–1060.
Horn, J.W., van Ee, B.W., Morawetz, J.J., Riina, R., Steinmann, V.W., Berry, P.E., and Wurdack, K.J. (2012). Phylogenetics and the evolution of major structural characters in the giant genus Euphorbia L. (Euphorbiaceae). Molecular Phylogenetics and Evolution **63**: 305–326.
Horn, J.W., Xi, Z., Riina, R., Peirson, J.A., Yang, Y., Dorsey, B.L., Berry, P.E., Davis, C.C., and Wurdack, K.J. (2014). Evolutionary bursts in *Euphorbia* (Euphorbiaceae) are linked with photosynthetic pathway: EVOLUTIONARY BURSTS IN *EUPHORBIA*. Evolution **68**: 3485–3504.
Hua, J., Liu, Y., Xiao, C.-J., Jing, S.-X., Luo, S.-H., and Li, S.-H. (2017). Chemical profile and defensive function of the latex of Euphorbia peplus. Phytochemistry **136**: 56–64.
Jiang, S.-Y., Jin, J., Sarojam, R., and Ramachandran, S. (2019). A Comprehensive Survey on the Terpene Synthase Gene Family Provides New Insight into Its Evolutionary Patterns. Genome Biology and Evolution **11**: 2078–2098.

Jones, C.G., Moniodis, J., Zulak, K.G., Scaffidi, A., Plummer, J.A., Ghisalberti, E.L., Barbour, E.L., and Bohlmann, J. (2011). Sandalwood Fragrance Biosynthesis Involves Sesquiterpene Synthases of Both the Terpene Synthase (TPS)-a and TPS-b Subfamilies, including Santalene Synthases. Journal of Biological Chemistry **286**: 17445–17454.
Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods **14**: 587–589.
Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research **30**: 3059–3066.
Kearse, M. et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28**: 1647–1649.
Konno, K. (2011). Plant latex and other exudates as plant defense systems: Roles of various defense chemicals and proteins contained therein. Phytochemistry **72**: 1510–1530.
Kubitzki, K. and Gottlieb, O.R. (1984). PHYTOCHEMICAL ASPECTS OF ANGIOSPERM ORIGIN AND EVOLUTION. Acta Botanica Neerlandica **33**: 457–468.
Lazreg Aref, H., Mosbah, H., Fekih, A., and Kenani, A. (2014). Purification and Biochemical Characterization of Lipase from Tunisian Euphorbia peplus Latex. J Am Oil Chem Soc **91**: 943–951.
Lewinsohn, T.M. (1991). The geographical distribution of plant latex. Chemoecology **2**: 64–68.
Lin, J., Wang, D., Chen, X., Köllner, T.G., Mazarei, M., Guo, H., Pantalone, V.R., Arelli, P., Stewart, C.N., Wang, N., and Chen, F. (2017). An (*E,E*)-α-farnesene synthase gene of soybean has a role in defence against nematodes and is involved in synthesizing insect-induced volatiles. Plant Biotechnol J **15**: 510–519.
Loureiro, J., Rodriguez, E., Dolezel, J., and Santos, C. (2007). Two New Nuclear Isolation Buffers for Plant DNA Flow Cytometry: A Test with 37 Species. Annals of Botany **100**: 875–888.
Luo, D., Callari, R., Hamberger, B., Wubshet, S.G., Nielsen, M.T., Andersen-Ranberg, J., Hallström, B.M., Cozzi, F., Heider, H., Lindberg Møller, B., Staerk, D., and Hamberger, B. (2016). Oxidation and cyclization of casbene in the biosynthesis of *Euphorbia* factors from mature seeds of *Euphorbia lathyris* L. Proc Natl Acad Sci USA **113**: E5082–E5089.
Maeda, H.A. (2019). Evolutionary Diversification of Primary Metabolism and Its Contribution to Plant Chemical Diversity. Front. Plant Sci. **10**: 881.
Maghuly, F., Vollmann, J., and Laimer, M. (2015). Biotechnology of Euphorbiaceae (Jatropha curcas, Manihot esculenta, Ricinus communis). In Applied Plant Genomics and Biotechnology (Elsevier), pp. 87–114.
Martin, D.M., Aubourg, S., Schouwey, M.B., Daviet, L., Schalk, M., Toub, O., Lund, S.T., and Bohlmann, J. (2010). Functional Annotation, Genome Organization and Phylogeny of the Grapevine (Vitis vinifera) Terpene Synthase Gene Family Based on Genome Assembly, FLcDNA Cloning, and Enzyme Assays. BMC Plant Biol **10**: 226.
Mergner, J., Frejno, M., Messerer, M., Lang, D., Samaras, P., Wilhelm, M., Mayer, K.F.X., Schwechheimer, C., and Kuster, B. (2020). Proteomic and transcriptomic profiling of aerial organ development in Arabidopsis. Sci Data **7**: 334.
Michaelson, M.J., Price, H.J., Ellison, J.R., and Johnston, J.S. (1991). COMPARISON OF PLANT DNA CONTENTS DETERMINED BY FEULGEN

MICROSPECTROPHOTOMETRY AND LASER FLOW CYTOMETRY. American Journal of Botany **78**: 183–188.
Minh, B.Q., Nguyen, M.A.T., and von Haeseler, A. (2013). Ultrafast Approximation for Phylogenetic Bootstrap. Molecular Biology and Evolution **30**: 1188–1195.
Nemethy, E., Skrukrud, C., Piazza, G., and Calvin, M. (1983). Terpenoid biosynthesis in Euphorbia latex. Biochimica et Biophysica Acta (BBA) - General Subjects **760**: 343–349.
Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution **32**: 268–274.
Occdownload Gbif.Org (2020). Occurrence Download.: 925941.
Ouyang, S. et al. (2007). The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Research **35**: D883–D887.
Parr, C.S., Wilson, N., Leary, P., Schulz, K., Lans, K., Walley, L., Hammock, J., Goddard, A., Rice, J., Studer, M., Holmes, J., and Corrigan, Jr., R. (2014). The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. BDJ **2**: e1079.
Pava, M.D.P.R., Darly Gabriela Muñoz Lara, Camayo, M.A.R., Trujillo, L.F.F., Francy Andrea Muñoz Castro, and Muñoz, N.P. (2017). Colección Mastozoológica del Museo de Historia Natural de la Universidad del Cauca.: 251 records.
Pellicer, J. and Leitch, I.J. (2014). The Application of Flow Cytometry for Estimating Genome Size and Ploidy Level in Plants. In Molecular Plant Taxonomy, P. Besse, ed, Methods in Molecular Biology. (Humana Press: Totowa, NJ), pp. 279–307.
Perry, B.A. (1943). CHROMOSOME NUMBER AND PHYLOGENETIC RELATIONSHIPS IN THE EUPHORB ACEAE. American Journal of Botany **30**: 527–543.
Pichersky, E. and Lewinsohn, E. (2011). Convergent Evolution in Plant Specialized Metabolism. Annu. Rev. Plant Biol. **62**: 549–566.
Prenner, G. and Rudall, P.J. (2007). Comparative ontogeny of the cyathium in Euphorbia (Euphorbiaceae) and its allies: exploring the organ-flower-inflorescence boundary. Am. J. Bot. **94**: 1612–1629.
R Studio Team (2020). RStudio: Integrated development for R (RStudio, Inc.).
Ramos, M.V., Demarco, D., da Costa Souza, I.C., and de Freitas, C.D.T. (2019). Laticifers, Latex, and Their Role in Plant Defense. Trends in Plant Science **24**: 553–567.
Ramos, M.V., Freitas, C.D.T., Morais, F.S., Prado, E., Medina, M.C., and Demarco, D. (2020). Plant latex and latex-borne defense. In Advances in Botanical Research (Elsevier), pp. 1–25.
Rhee, S.Y. (2003). The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Research **31**: 224–228.
Rudall, P. (1994). Laticifers in Crotonoideae (Euphorbiaceae): Homology and Evolution. Annals of the Missouri Botanical Garden **81**: 270.
Salatino, A., Salatino, M.L.F., and Negri, G. (2007). Traditional uses, chemistry and pharmacology of Croton species (Euphorbiaceae). J. Braz. Chem. Soc. **18**: 11–33.
Salehi et al. (2019). Euphorbia-Derived Natural Products with Potential for Use in Health Maintenance. Biomolecules **9**: 337.

Salem, M., Bernach, M., Bajdzienko, K., and Giavalisco, P. (2017). A Simple Fractionated Extraction Method for the Comprehensive Analysis of Metabolites, Lipids, and Proteins from a Single Sample. JoVE: 55802.
Samuel, R., Kathriarachchi, H., Hoffmann, P., Barfuss, M.H.J., Wurdack, K.J., Davis, C.C., and Chase, M.W. (2005). Molecular phylogenetics of Phyllanthaceae: evidence from plastid MATK and nuclear PHYC sequences. Am. J. Bot. **92**: 132–141.
Santana, K.C.B., Pinangé, D.S.B., Vasconcelos, S., Oliveira, A.R., Brasileiro-Vidal, A.C., Alves, M.V., and Benko-Iseppon, A.M. (2016). Unraveling the karyotype structure of the spurges Euphorbia hirta Linnaeus, 1753 and E. hyssopifolia Linnaeus, 1753 (Euphorbiaceae) using genome size estimation and heterochromatin differentiation. CCG **10**: 657–696.
Scientific Data Curation Team (2020). Metadata record for: Proteomic and transcriptomic profiling of aerial organ development in Arabidopsis.: 4824 Bytes.
Sharma, N., Samarakoon, K., Gyawali, R., Park, Y.-H., Lee, S.-J., Oh, S., Lee, T.-H., and Jeong, D. (2014). Evaluation of the Antioxidant, Anti-Inflammatory, and Anticancer Activities of Euphorbia hirta Ethanolic Extract. Molecules **19**: 14567–14581.
Shu-Cherng, F. and Tsao, H.-S.J. (2001). Entropy Optimization: Shannon Measure of Entropy and its Properties. In Encyclopedia of Optimization, C.A. Floudas and P.M. Pardalos, eds (Springer US: Boston, MA), pp. 552–558.
de Souza, L.S., Puziol, L.C., Tosta, C.L., Bittencourt, M.L.F., Ardisson, J.S., Kitagawa, R.R., Filgueiras, P.R., and Kuster, R.M. (2019). Analytical methods to access the chemical composition of an Euphorbia tirucalli anticancer latex from traditional Brazilian medicine. Journal of Ethnopharmacology **237**: 255–265.
Spanò, D., Pintus, F., Mascia, C., Scorciapino, M.A., Casu, M., Floris, G., and Medda, R. (2012). Extraction and characterization of a natural rubber from Euphorbia characias latex. Biopolymers **97**: 589–594.
Steckel, A. and Schlosser, G. (2019). An Organic Chemist's Guide to Electrospray Mass Spectrometric Structure Elucidation. Molecules **24**: 611.
Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., and Arita, M. (2015). MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods **12**: 523–526.
Van Parijs, J., Broekaert, W.F., Goldstein, I.J., and Peumans, W.J. (1991). Hevein: an antifungal protein from rubber-tree (Hevea brasiliensis) latex. Planta **183**: 258–264.
Verma, V.N. (2013). The Chemical Study of *Calotropis*. ILCPA **20**: 74–90.
Webster, G.L. (1994). Classification of the Euphorbiaceae. Annals of the Missouri Botanical Garden **81**: 3.
Webster, G.L., Brown, W.V., and Smith, B.N. (1975). SYSTEMATICS OF PHOTOSYNTHETIC CARBON FIXATION PATHWAYS IN EUPHORBIA. TAXON **24**: 27–33.
Wickham, H. (2009). Ggplot2: elegant graphics for data analysis (Springer: New York).
Wiens, J. and Hernandez, T. (2019). Why are there so many flowering plants? A multi-scale analysis of plant diversification.: 1800831 bytes.
Wink, M. (2020). Evolution of the Angiosperms and Co-evolution of Secondary Metabolites, Especially of Alkaloids. In Co-Evolution of Secondary Metabolites, J.-M. Mérillon and K.G. Ramawat, eds, Reference Series in Phytochemistry. (Springer International Publishing: Cham), pp. 151–174.

Wu, Y., Zhou, Y., Saveriades, G., Agaian, S., Noonan, J.P., and Natarajan, P. (2013). Local Shannon entropy measure with statistical tests for image randomness. Information Sciences **222**: 323–342.
Yu, J., Xue, J.-H., and Zhou, S.-L. (2011). New universal matK primers for DNA barcoding angiosperms. Journal of Systematics and Evolution **49**: 176–181.
Yu, Z., Zhao, C., Zhang, G., Teixeira da Silva, J.A., and Duan, J. (2020). Genome-Wide Identification and Expression Profile of TPS Gene Family in Dendrobium officinale and the Role of DoTPS10 in Linalool Biosynthesis. IJMS **21**: 5419.
Züst, T., Petschenka, G., Hastings, A.P., and Agrawal, A.A. (2019). Toxicity of Milkweed Leaves and Latex: Chromatographic Quantification Versus Biological Activity of Cardenolides in 16 Asclepias Species. J Chem Ecol **45**: 50–60.

SUPPLEMENT

Part 1. MS-DIAL Parameters (Section 2.2)

For the Analysis Parameter setting, 0.005 Da and 0.01 Da were used for MS1 and MS2 tolerances respectively. The minimum peak height was set to 1000 amplitude and the mass slice width was set to 0.1 Da. The sigma window value was set to 0.5 and MS/MS abundance cut off was set to 0 amplitude. For the identification set up, retention time was set to 100 min, accurate m mass tolerance (MS1) and (MS2) were set to 0.01 and 0.05 Da respectively, and identification score cut off was set to 80%. Three different adducts were selected for this positive alignment file: [M+H]+, [M+NH4]+, and [M+Na]+. For the alignment section, the result name was set to alignmentResult_2021_3_7_23_30_44, the reference file was set to 20190429_EU_SJ_02, retention time tolerance was 0.1 min, and MS1 tolerance was 0.015 Da. Under the advanced option section, “Remove features based on blank information” was selected. Lastly, Isotope tracking was left as the default setting.

Part 2. MS-DIAL Parameters (Section 2.3)

For the Analysis Parameter setting, 0.005 Da and 0.01 Da were used for MS1 and MS2 tolerances respectively. The minimum peak height was set to 10000 amplitude and the mass slice width was set to 0.05 Da. The sigma window value was set to 0.5 and MS/MS abundance cut off was set to 0 amplitude. For the identification set up, retention time was set to 100 min, accurate mass tolerance (MS1) and (MS2) were set to 0.01 and 0.05 Da respectively, and identification score cut off was set to 80%. Five different adducts were selected for this positive alignment file: [M+H]+, [M+NH4]+, [M+Na]+, [M+ACN+H]+, and [2M+H]+. For the alignment section, the result name was set to alignmentResult_2021_2_28_14_33_48, the reference file was set to 210120_Peplus_1, retention time tolerance was 0.05 min, and MS1 tolerance was 0.015 Da. Under the advanced option section, "Remove features based on blank information" was selected. Lastly, Isotope tracking was left as the default setting.

Part 3. Hmmsearch command lines (Section 2.5)

hmmsearch
--cut_ga/Users/bagsejin/Desktop/Fall2020/Courses/Research/HMM/C.hmm
/Users/bagsejin/Desktop/Fall2020/Courses/Research/FolderName/PlantName.fasta >
/Users/bagsejin/Desktop/Fall2020/Courses/Research/FolderName/PantNameOut.txt

hmmsearch
--cut_ga /Users/bagsejin/Desktop/Fall2020/Courses/Research/HMM/N.hmm
/Users/bagsejin/Desktop/Fall2020/Courses/Research/FolderName/PlantName.fasta >
/Users/bagsejin/Desktop/Fall2020/Courses/Research/FolderName/PantNameOut.txt

Part 4. Sample Genome Size Calculations (Sections 3.3 and 4.1)

Solanum lycopersicum Genome Size

- Trial 1
  - $\frac{mean\ position\ of\ Tomato}{Mean\ position\ of\ CRBC} * 2.33\ pg = \frac{28303}{19830} * 2.33 = \mathbf{3.33\ pg}$
- Trial 2
  - $\frac{mean\ position\ of\ Tomato}{Mean\ position\ of\ CRBC} * 2.33\ pg = \frac{23351}{19830} * 2.33 = \mathbf{2.74\ pg}$
- Trial 3
  - $\frac{mean\ position\ of\ Tomato}{Mean\ position\ of\ CRBC} * 2.33\ pg = \frac{27076}{19830} * 2.33 = \mathbf{3.18\ pg}$
- Average: 3.08 pg

E. peplus Genome Size

- Trial 1
  - $\frac{mean\ position\ of\ peplus}{Mean\ position\ of\ CRBC} * 2.33\ pg = \frac{11215}{19830} * 2.33 = \mathbf{1.32\ pg}$
- Trial 2
  - $\frac{mean\ position\ of\ peplus}{Mean\ position\ of\ CRBC} * 2.33\ pg = \frac{9100}{19830} * 2.33 = \mathbf{1.07\ pg}$
- Trial 3
  - $\frac{mean\ position\ of\ peplus}{Mean\ position\ of\ CRBC} * 2.33\ pg = \frac{8571}{19830} * 2.33 = \mathbf{1.01\ pg}$
- Average: 1.13 pg

E. lathyris Genome Size

- Trial 1
  - $\frac{mean\ position\ of\ lathyris}{Mean\ position\ of\ CRBC} * 2.33\ pg = \frac{53260}{19830} * 2.33 = \mathbf{6.26\ pg}$
- Trial 2
  - $\frac{mean\ position\ of\ lathyris}{Mean\ position\ of\ CRBC} * 2.33\ pg = \frac{30007}{19830} * 2.33 = \mathbf{3.53\ pg}$
- Trial 3
  - $\frac{mean\ position\ of\ lathyris}{Mean\ position\ of\ CRBC} * 2.33\ pg = \frac{31603}{19830} * 2.33 = \mathbf{3.71\ pg}$
- Average: 4.50 pg