Building maps of plant surface chemistry using literature and citizen-collected mass spectrometry samples

A Thesis

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Dien Nguyen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE IN CHEMISTRY

Advisor: Lucas Busta

June 2022

Abstract

Understanding plant chemicals can greatly help achieve the important goals of reducing negative environmental impacts of agriculture and achieving high crop yields. One class of chemicals that has a big influence on plant health and growth is triterpenoid. However, there is a limited number of biological systems in which to study triterpenoids in detail. To increase the number of study systems, and thus our potential to build knowledge of triterpenoid function, there is a critical need to understand which triterpenoids can be found on which plant species' surface. The objective of this project was to build maps of plant surface compound presence with an emphasis on triterpenoids, using published and newly acquired gas chromatography-mass spectrometry data. The maps reveal trends that some triterpenoids are very commonly found on plant surface, while others are rare. In addition, divergent and convergent evolution traits among plants regarding their triterpenoid and other wax compounds presence were also identified.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

B.P.	before present	1
HYV	high yield varieties	5
DCAM	methyl dihydrocanaric acid	25
DRAM	methyl-3,4-seco-urs-12-en-3-oate	25
DLAM	methyl-4(23)-dihydro-lacunosic acid	25
DNAM	methyl-4(23)-dihydro-nyctanthic acid	25
GC-MS	gas chromatography-mass spectrometry	43
BSTFA	N,O-Bis(trimethylsilyl)trifluoroacetamide	44
TIC	total ion chromatogram	44
GC-TCD	gas chromatography-thermal conductivity detector	47

Chapter 1: General Introduction

Plant cultivation has been a crucial and inseparable part of human history because plants are a major source of food. Agriculture helped to transform human society from the hunting-and-gathering lifestyle into the civilizations today, via a period known as the Neolithic Revolution or the First Agricultural Revolution. As hunting and gathering has been a primary and efficient way of life since the emergence of *homo sapiens* 200,000 years ago, there has been a lot of debate on why such transformation took place. A commonly accepted theory is due to the climate change that occurred at the end of the Ice Age, when the weather condition favors annual plants that die off during the long dry season, leaving dormant seed or tubers. The abundance of grain products that were also suitable for long term storage in addition to population pressure allowed humans to move toward a lifestyle of farming settlement (Larson et al. 2014). This transformation allowed the human population to grow many times larger, sustained by crop food, and would not be possible for nomadic hunters-gatherers (Bocquet-Appel 2011).

Exact dates were difficult to identify, as humans collected and consumed grains and vegetable products long before domesticating them. Studies have shown that agriculture started developing independently in different parts of the globe (see Figure 1) from 12,000-4200 before present (B.P.), and this included a very wide range of different plants and animals from different periods (see Figure 2). These regions were the sources of the major modern domesticates that spread to the adjacent area (Larson et al. 2014).

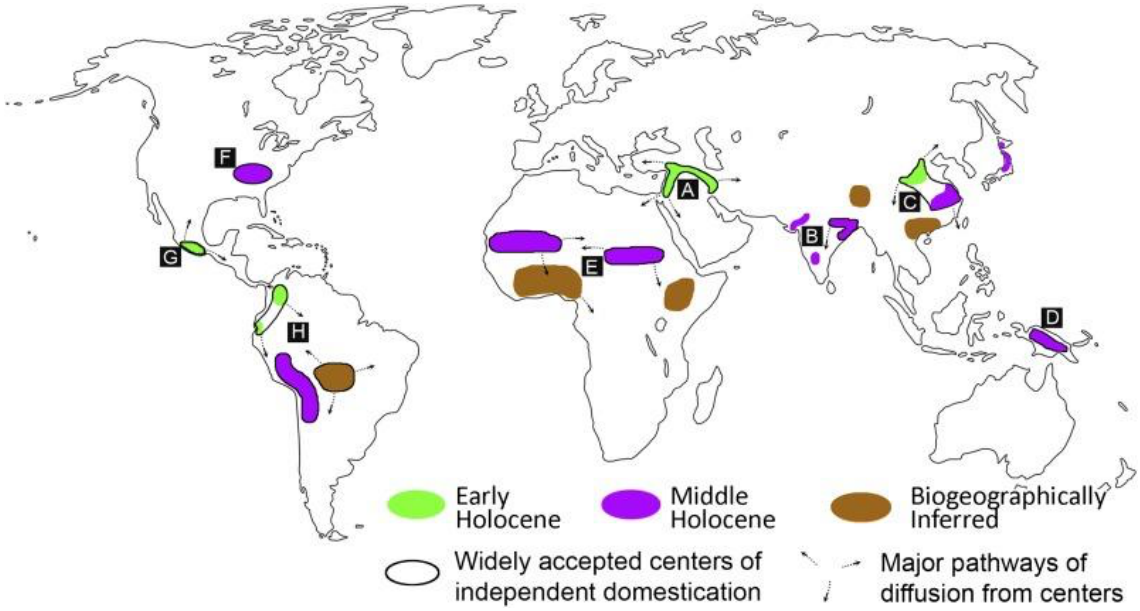

Figure 1 (Larson et al. 2014). A map showing likely domestication centers of at least one plant or animal with arrows indicate major diffusion of domesticates. The green region represents the domestication process during the early Holocene (12,000-8,200 B.P.) and the purple region presents domestication process during the middle Holocene (8,200-4,200 B.P.); while brown regions represent domestication with only evidence based on the presence of domestic forms indigenous to the region outside of their native distribution (unclear domestication time). Letters A-H correspond to the chart in Figure 2.

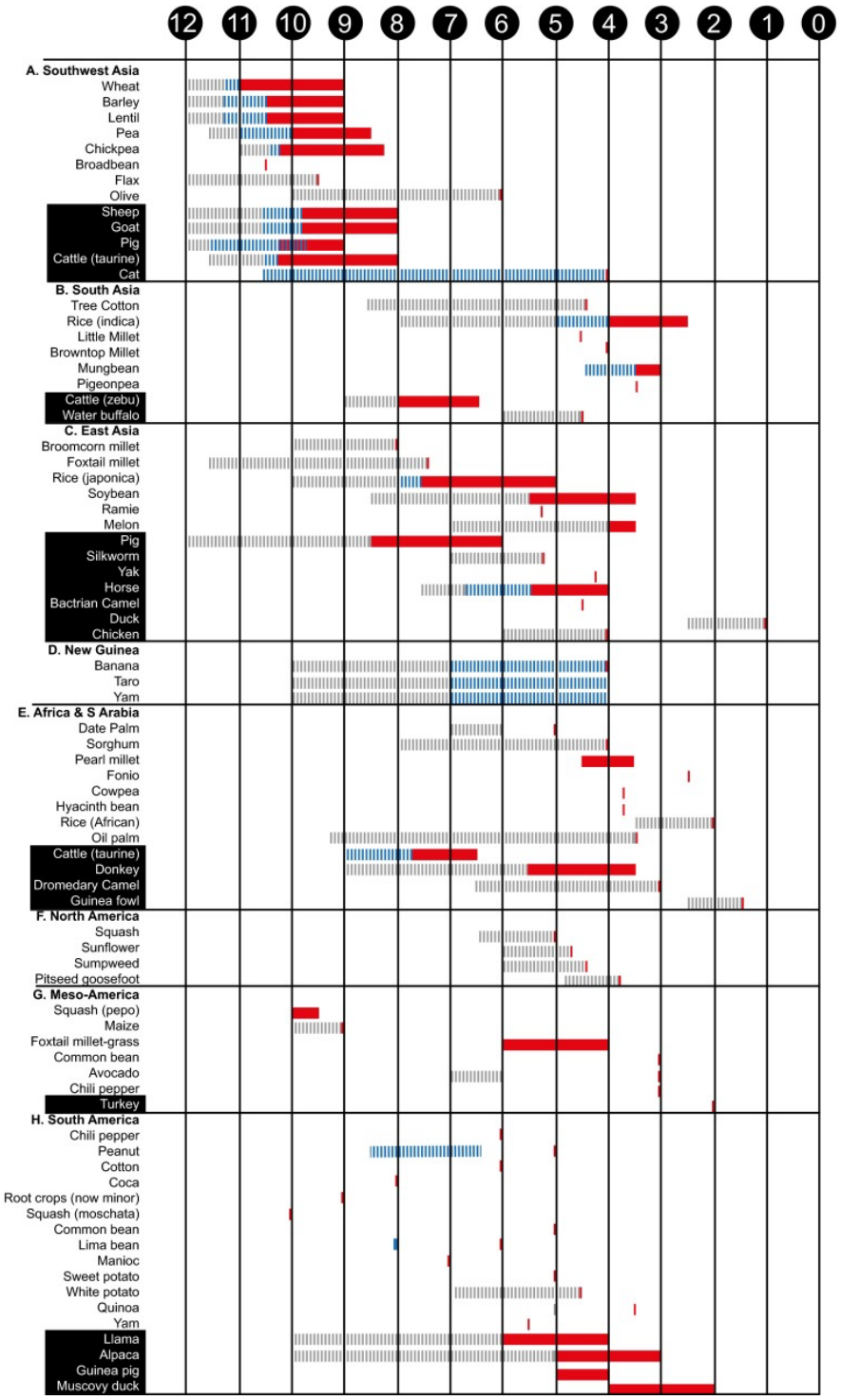

Figure 2 (Larson et al. 2014). A chronological chart with regions and time frames that shows where and when key plants and animals were domesticated. The numbers in black circle represent thousands of years B.P. Gray dashed lines represent documented exploitation before domestication; blue dashed lines represent pre-domestication cultivation; and most importantly, red bars represent actual domestication with evidence of morphological changes.

During the Middle Ages to the Early Modern Period, the human population continued to grow with the rapid increase in land cultivation. Trade and cultural exchange allowed for local farming techniques and different plant products to spread around the world. A significant event in this period is the Columbian exchange during the late 15th-18th century: an exchange of agricultural products between the Old and the New World. Maize, potatoes, sweet potatoes, and cassava moved from the New World to the Old World; meanwhile wheat, rice and barley moved in the opposite direction. Maize and cassava became the main staple foods in Africa (Wambugu et al. 2000) and replaced the native crops; and the South American potatoes became a staple crop in Europe in addition to the available wheat (Nunn et al. 2011). The movement of crops in both directions of the Atlantic Ocean increased food supply, thus increased births and reduced mortality, and caused a population boom that left a lasting effect in many cultures around the world during this period. For example, it is estimated that potatoes contributed to 25% of population growth in Afro-Eurasia between 1700-1900s (Nunn et al. 2011) and the South American tomatoes became an integral part of Italian cuisine and culture (Smith 1994).

Another significant event happening at the end of the Columbian exchange was the British Agricultural Revolution around the 17th-19th century. New practices include enclosure, advancement in irrigation and fertilizers, mechanization and selective breeding helped to increase agricultural output with groundbreaking efficiency. This led to unprecedented population growth and also freed up a lot of the workforce to drive the concurring Industrial Revolution (MacDonald 2014). Unfortunately, famines were still a significant problem in the 20th century with millions of people who have died in each of

the ten major famines between 1920-1990s due to crop failure in combination with government policy and war. Between the 1940s and 1970s, a series of research and development initiatives known as the Green Revolution took place. These initiatives were pioneered by agricultural scientist Norman Borlaug, a University of Minnesota graduate. He received the Nobel Peace Prize in 1970, is considered to be “the Father of the Green Revolution" and credited in saving over a billion people from starvation (Hazell 2009).

The Green Revolution aimed at two main biochemistry technological areas: cultivation and breeding. Cultivation research focused on improvement in irrigation projects, agricultural machinery, pesticides and especially with the use of synthetic fertilizers (Levetin et al. 2020). Synthetic nitrogen fertilizers can be seen to have made a major contribution to the dramatic increase of agricultural output. This was demonstrated by the global population that can be supported with and without the use of synthetic nitrogen fertilizers in crops in Figure 3 (Ritchie et al. 2013). The breeding area focuses on combining selective breeding with modern genetics knowledge for development of high yield varieties (HYV) crops including rice, maize, and wheat. Even though traditional crops can survive better than these HYVs in normal conditions, HYVs outperform traditional crops in the presence of cultivation conditions like irrigation, pesticides, and fertilizers.

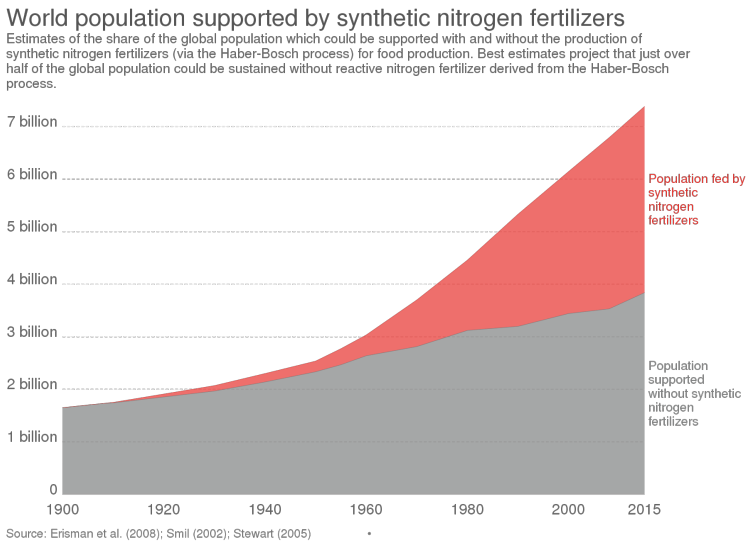

Figure 3 (Ritchie et al. 2013). Estimated world population supported by agriculture production with or without synthetic nitrogen fertilizers over time.

In 1789, Thomas Malthus famously warned that the Earth would not be able to support the growing human population. Malthusian Paul Ehrlich said in his book that “India couldn't possibly feed two hundred more million people by 1980", though this failed to happen with the introduction of Borlaug's wheat HYVs. Between 1950 and 1985, global grain production increased by 160% and the Green Revolution technologies helped to produce food for the recently growing population and avoid widespread famine. However, Borlaug understood the reality of human overpopulation and said in his Nobel lecture: “The Green Revolution has won a temporary success in man's war against hunger and deprivations … But the frightening power of human reproduction must also be curbed; otherwise the success of the green revolution will be ephemeral only.” The Green Revolution has greatly increased crop yield, but the yield potential has remained steady in the recent decades due to some factors like the emergence of herbicide-resistant weeds within two decades and insecticide-resistant insects within a decade. This proves the need for constant improvement in agricultural development through plant chemistry.

Crop yields, particularly under stressful conditions like those created by climate change, are linked to a crop plant's resilience. Plant's resilience in turn, like that of other organisms, is greatly influenced by chemistry: understanding plant chemistry offers great strategies for crop protection and food productivity. In general, there are three aspects in which plant chemistry can contribute to improvement in agricultural production and usage (Smith et al. 2008). First, chemical experiments can advance knowledge in new molecules and information on environmental and toxicological properties as well as biological activities of plants. Secondly, chemistry can push for advancements in synthetic compounds that can be manufactured and used for crop protection more efficiently and more environmentally friendly. And finally, research in non-food crop research can also advance the potential of agriculture to provide fuels from renewable resources. Given these reasons, an advancement in plant chemistry is undeniably an advancement in agricultural potential.

Understanding plant chemicals can therefore help meet the important goals of efficient plant growth and high crop yields. These goals in turn will help to (i) reduce the negative environmental impacts of agriculture and (ii) produce food using fewer resources.

One class of chemicals that has a big influence on plant health and growth is the triterpenoids. These compounds have a general molecular formula of $C_{30}H_nO_m$, usually have five 5- or 6-member carbons rings, as well as one or more oxygen-containing functional groups (see Figure 4). They also have a variety of bioactivities. For example,

triterpenoids like lupeol and ursolic acid are associated with post-harvest weight and firmness of highbush blueberries (*Vaccinium corymbosum*) (Moggia et al. 2016), the triterpenoids α-amyrin, β-amyrin, and simiarenol seem to help the grain crop *Sorghum bicolor* be exceptionally tolerant to drought conditions (Busta et al. 2021), and the triterpenoids ursolic acid, echinocystic acid, and oleanolic acid help the desert shrub *Rhazya stricta* stay hydrated at elevated temperature (Schuster et al 2016).

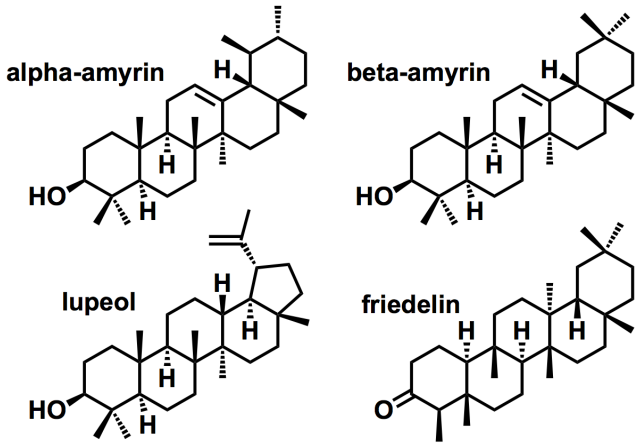

Figure 4. Some common and structurally simple triterpenoids.

Triterpenoids can also help defend a plant against other living organisms. For example, birch trees produce the triterpenoid betulin that deters the Colorado Potato beetle (Huang et al., 1995), cork trees produce the triterpenoid friedelane to defend against the parasites *Trypanosoma cruzi* and *Leishmania infantum* (Moiteiro et al., 2006), and citrus trees produce limonoid triterpenoids with anti-malarial (against *Plasmodium falciparum*), anti-microbial (against bacteria *Streptococcus pyogenes*, *Staphylococcus aureus,* and *Botrytis cinerea*) and insecticidal (against fall armyworm) activities (Roy et al., 2006).

Finally, triterpenoids also serve as building blocks for more structurally complex classes of biological defense compounds. One example of such triterpenoid derivatives is saponins. Saponins are amphipathic compounds made from a combination of fat-soluble triterpenoid rings and a water-soluble oligosaccharide chain (see Figure 5). Experimental evidence suggests that their amphipathic character allows saponins to permeabilize membranes, particularly the intestinal cells of insects such as pea aphid and cotton leafworm (De Geyter et al., 2011; De Geyter 2012), leading to feeding deterrence, growth inhibition, or outright fatality (Singh et al. 2018). In agriculture, these pest management functions are usually carried out using synthetic pesticides, but synthetic pesticides are often not biodegradable and can be toxic to humans and other non-target organisms, causing damage to the environment and negatively affecting the ecosystem. Replacing synthetic pesticides with natural plant metabolites like saponins (triterpenoid derivatives), therefore, will likely achieve three important objectives: (i) ensure the safety of the crop, (ii) minimize the toxicity of crop products, and (iii) make agriculture more sustainable (Singh et al. 2018).

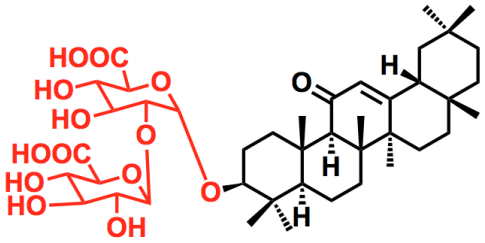

Figure 5: An example of the amphipathic saponin group compounds: glycyrrhizin. This compound includes a nonpolar triterpenoid group (black) and a polar & water-soluble sugar group (red).

Based on the studies summarized above, triterpenoids and their derivatives have the potential to make a big impact on our agricultural system, but only if we can understand

how they function. However, we have a *problem* in that we know of only a limited number of biological systems in which to study triterpenoid function in detail. Therefore, there is a *critical need* to understand which triterpenoid compounds can be found in which plant species so we can identify new study systems for future triterpenoid research. This is a critical need to which an analytical chemist can make a major contribution. Accordingly, the *objective of the work* described in this thesis is to build "maps" (example in Figure 6) that show the presence of different surface wax compounds in plants with emphasis on triterpenoids along with their relative abundance using (i) existing, published data and (ii) newly generated data.

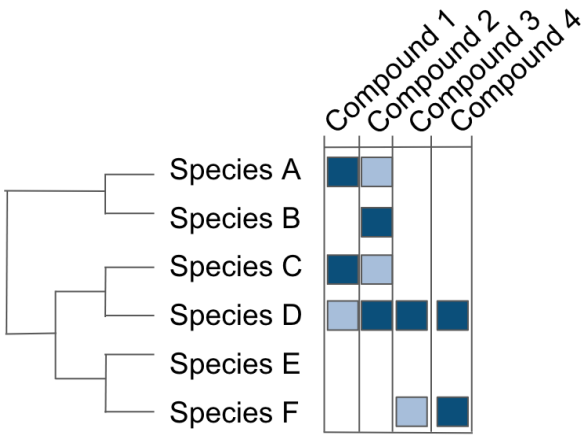

Figure 6: An example of a "map" of compound occurrence. a phylogenetic tree of all the plant species on the left; a table indicating which triterpenoid compound is absent (white), minor (light blue) or major (dark blue) within wax tissues in each respective species.

In order to meet the objective of this proposal, two aims will be carried out and described in the two following chapters:

Aim 1: Document known occurrences of plant triterpenoids based on published literature studies.

Aim 2: Document new occurrences of plant triterpenoids + other surface wax compounds using more than 100 samples.

Triterpenoids have been documented before on a species-by-species basis. In contrast, this study will focus on (i) systematically gathering and organizing information on existing sources of plant triterpenoids into a literature review and (ii) acquiring new triterpenoid occurrence data by analyzing samples coming from various plant species sent by citizen scientists. Each of these focus areas will be addressed in two respective chapters. The combined dataset generated during this project will be a collection of triterpenoid information on a large, plant kingdom-spanning scale, which has never been done on this scale before. In addition to that, the second chapter will also document other plant wax compounds alongside triterpenoids and will explore how these compounds interact or have any influences on the triterpenoids being produced at all. The maps will give us a greater number of biological systems in which to study triterpenoids and triterpenoid derivatives like saponins. This will greatly contribute to future experiments that directly test triterpenoid function. This in turn will help to make agricultural production better in three different ways: (i) keep crops safe, (ii) reduce pesticides residues on plant products, and (iii) make agriculture more environmentally friendly.

Chapter 2: Meta-analysis

1. Introduction

Many published articles document the presence of triterpenoids in specific plant species; their importance in plant biology as well as their potential in agriculture (see Chapter 1) and their role in medicines (Han et al. 2015). Triterpenoids can have different structures (although each has a general molecular formula of $C_{30}H_nO_m$) and are found in aerial surfaces of a variety of plants in different amounts. They have been documented in numerous literature sources as being able to reduce water loss and serve in bio-defense roles (see Chapter 1). On the other hand, these triterpenoid documentations are fragmented from different sources and many articles simply documented the existence of triterpenoids on a given plant species without much further analysis of their presence. On top of that, the plants containing these compounds have not been analyzed in a way that triterpenoids can be observed through a phylogenetic relationship from a variety of plant species. Because of these limitations, there is an opportunity to collect triterpenoid information from these various articles and summarize them on a more substantial scale and with in-depth analysis. If this is to be realized, in the short term it can help people understand plant triterpenoids evolution, and in the long term, it will help to provide the foundation for future experiments to understand triterpenoids and to improve their potential in activities like agriculture (see Chapter 1).

The information in this chapter comes from published scientific articles, in contrast to experimental information that will appear in chapter two. The objective in this chapter is to build two triterpenoid maps on a sizable scale which was never done before, and thus allow for the analysis of triterpenoids in plants and conclusion about the observed trends from many different angles. There are several searching tools that can be used including Google Scholar, SciFinder, Science Direct. On one hand, Science Direct searches are limited to only what the publishers publish and SciFinder does not provide direct links to the articles, so DOI or Google Scholar searches are needed after. In contrast, Google Scholar shows the library access with links for each article that comes up in the search; thus, this makes searching and going through each article a lot faster and more efficient. Therefore, Google Scholar was chosen as it permits quick and easy access in a non-selective fashion to a wide variety of articles that includes many open sources with general and simple queries. This chapter is essentially a meta-analysis chapter, where documented information is gathered and organized through multiple steps which leads into two maps. These comprehensive two maps include one map shown on the more general level of plant genus and one map on the more detailed species level. These maps will allow for analytical observation in different ways and interpretations in as much detail as possible about triterpenoid composition and function in the plant kingdom.

2. Methods

The goal of the work documented in this chapter was to construct two maps of triterpenoid occurrence based on published information. In order to build this map, three

major steps were performed: (i) finding and selecting scientific papers reporting the occurrence of specific triterpenoids from waxes of specific plant species, (ii) analyzing those papers and using them to build a spreadsheet of triterpenoid occurrence, and (iii) analyzing this spreadsheet and constructing an occurrence map.

2.1 Literature Search

The first step being performed was an extensive search for articles that mentioned specific triterpenoids occurring on the surfaces of specific plant species. Some examples of queries used in this search are: “triterpenoid”, “wax” and “plant”. Articles dated from the 1970s to as recently as 2020 were picked and from these, it was necessary to filter again and select articles with quantitative tables and with descriptions of instruments and extraction methods used, as well as the specific plant tissue studied. The resulting articles were organized into a spreadsheet with their authors, date, and title. These articles were then grouped into five categories: (i) articles describing triterpenoid occurrence, (ii) articles describing triterpenoid function, (iii) articles describing triterpenoid evolution, (iv) articles describing triterpenoid biosynthesis, and (v) titles to which full text access was not available. Then, as the aim is building a qualitative library of triterpenoid occurrence, only the papers that focused on triterpenoid occurrence were retained for more review.

2.2 Extract Data from Relevant Papers

With the list of triterpenoid occurrence articles in hand, it was time to start building the triterpenoid library. Each of these papers was gone through carefully and triterpenoids in each plant species were documented. To do this, eight things were recorded: (i) the plant genus, (ii) the plant species, (iii) the tissue location where it was being extracted, (iv) the extract solvent being used for sample analysis, (v) the instrumental system being used for analysis, (vi) the compound system/common name, (vii) their relative abundance in the sample (major or minor) and finally, (viii) their respective reference paper.

2.3 Analyze Extracted Literature Data

Using this literature data, it was possible to build two maps of triterpenoid occurrences, and this was done by utilizing the software RStudio along with the code included in **Appendix A**. All the plant species/genera were organized along their phylogenetic relationship while the all the unique triterpenoids are put into four groups based on the carbon structure of their five rings (complete ring indicated by five or six carbon in each ring or incomplete rings indicated by zero): 6-6-6-6-6, 6-6-6-6-5, 6-6-6-5-0 and 0-6-6-6-5. Each of these rings are marked with A, B, C, and D so the rings in the triterpenoid group can be observed through a similar ring position in each of them. As the maps were constructed, interpretation about trends of triterpenoids in plants can be documented on both the species and the genus level.

3. Results & Discussion

In order to understand and gather information of triterpenoid presence in plant species on a large scale, an extensive literature search and analysis of relevant articles was performed. This approach included a literature search using Google Scholar, data extraction based on each article's contents, plot creation with RStudio and trend identification.

The searches resulted in 51 different articles that were placed into five categories: (i) articles describing triterpenoid occurrence (37 articles), articles describing triterpenoid function (4 articles), articles describing triterpenoid evolution (4 articles), articles describing triterpenoid biosynthesis (2 articles), and titles to which full text access was not available (4 articles). Following the aim of building a qualitative library of triterpenoid occurrence, only the 37 papers in the first category were retained.

Each of these 37 papers were used to carefully record information in eight categories (see Method section 2.2 for more details). A problem about contradicting information from different sources about one type of triterpenoid did happen with seven different species from 11 papers. However, as these cases were examined carefully, it was found this happened since the maps created only report triterpenoid occurrences in relative terms, and most papers don't report all the same set of triterpenoids. For example, compound A is major and compound B is minor in paper 1 but compound B is major in comparison to another compound C in paper 2; so this is still true even though compound B is major in

one paper but is minor in another paper. This problem was later solved by making sure in such cases all triterpenoids set in the same species are being examined and their abundances are recorded in a consistent way. In the previous example, this means compound B should be recorded as a minor compound even though it is a lot more abundant than the very minor compound C. In the end, 696 triterpenoid occurrences (130 unique triterpenoids) were documented across 76 plant species from 41 genera based on 37 scientific articles.

Using the literature data of 696 triterpenoid occurrences, the two maps of triterpenoid occurrences (see Figure 7 and Figure 8) were built using the software RStudio. This was done on two main levels: the plant species level and the plant genus level. To show triterpenoid occurrence in a phylogenetic context, the maps contain four main components: a phylogenetic tree of all reported plant species or genera on the left, a heat map indicating whether each triterpenoid compound is minor or major in each respective species, a bar plot on top showing the number of species in which each triterpenoid has been found, and a bar plot on the right showing the number of triterpenoids each species has been reported to synthesize. On each level, the trending relationship can be analyzed in three different ways: the overall trend of the main components (heat map of all the species/genus + phylogenetic tree), the trend between the main components and the top plot, and the trends between the main component and the right plot.

As these analyses were being conducted, several conclusions can be drawn from the features of the two maps. First of all, (1) it seems clear that certain triterpenoids and

triterpenoids groups are present much more commonly in the documented genera/species, standing out with the rest of other triterpenoids/triterpenoid groups. Also, (2) documented plant genera/species that are phylogenetically close are more likely to have very similar triterpenoid profiles, even though (3) this trend does not necessarily correlate to the same triterpenoid diversity levels (total number of triterpenoid types within a genera/species). While this conclusion strongly reflects a divergent evolutionary trait from the documented plants, the maps also show (4) evidence of distantly related species that share multiple similar triterpenoids that likely happens from convergent evolution.

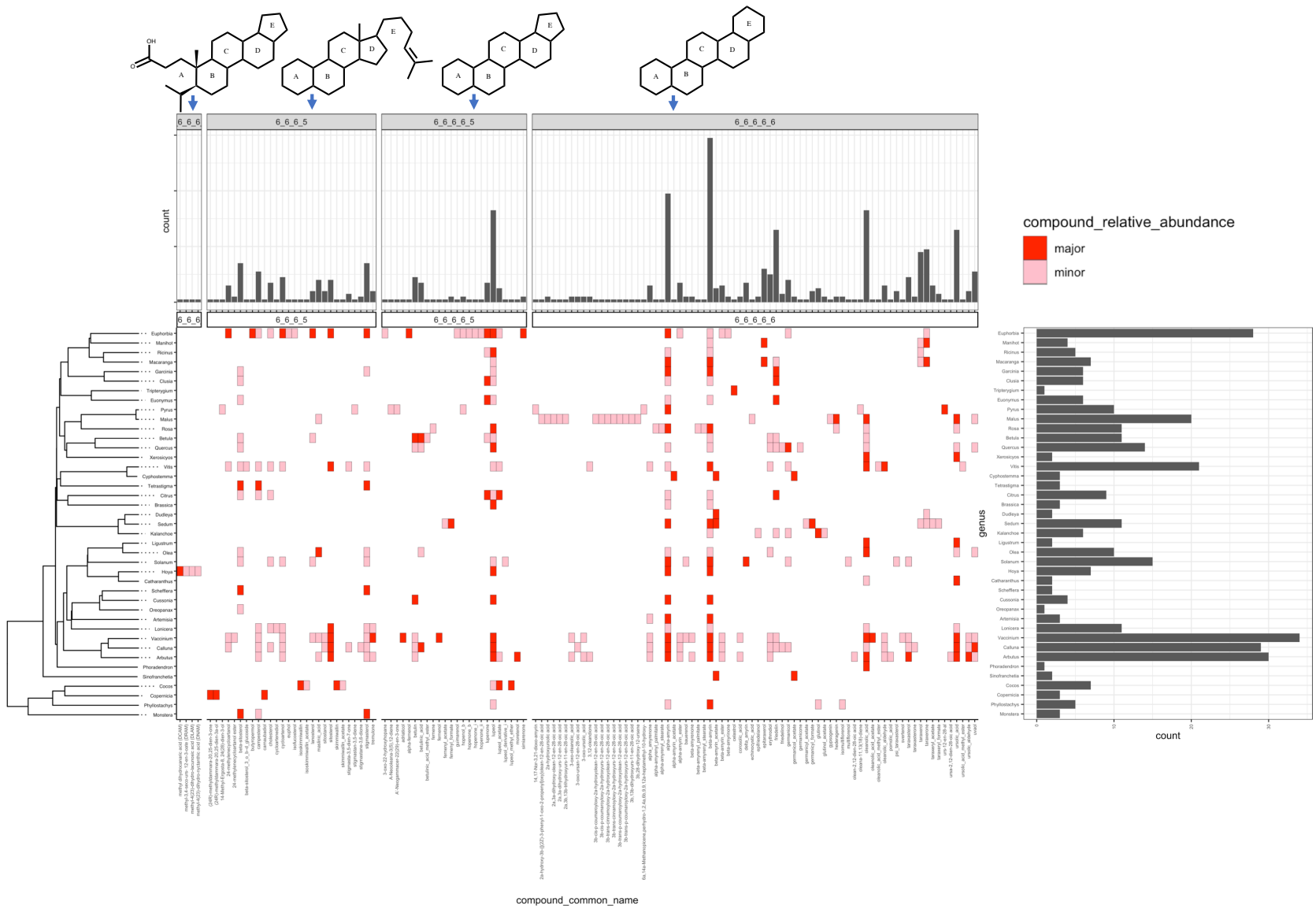
6_6_6
6_6_6_5
6_6_6_6_5
6_6_6_6_6
count
compound_relative_abundance
major
minor
genus
compound_common_name
count

Figure 7 (previous page). A comprehensive map of 696 triterpenoids occurrences across 41 different plant genera derived from literature information. On the heat map in the middle, the presence of triterpenoid occurrences is categorized into two components: dark red for the major compounds and light red for the minor compounds. The x-axis shows all the 130 unique triterpenoid names grouped into four groups based on the carbon structure of their five (complete or incomplete) rings: 6-6-6-6-6, 6-6-6-6-5, 6-6-6-5-0 and 0-6-6-6-5; and the bar plot on top demonstrates the number of occurrences of each of these triterpenoids across the studied plant genera. The y-axis shows all the plant genera grouped according to phylogenetic relationships (the relatedness of two adjacent genera on y-axis is indicated by the length of the branch between them); and the bar plot on the right shows the number of documented occurrences of unique triterpenoids in each of these genera.

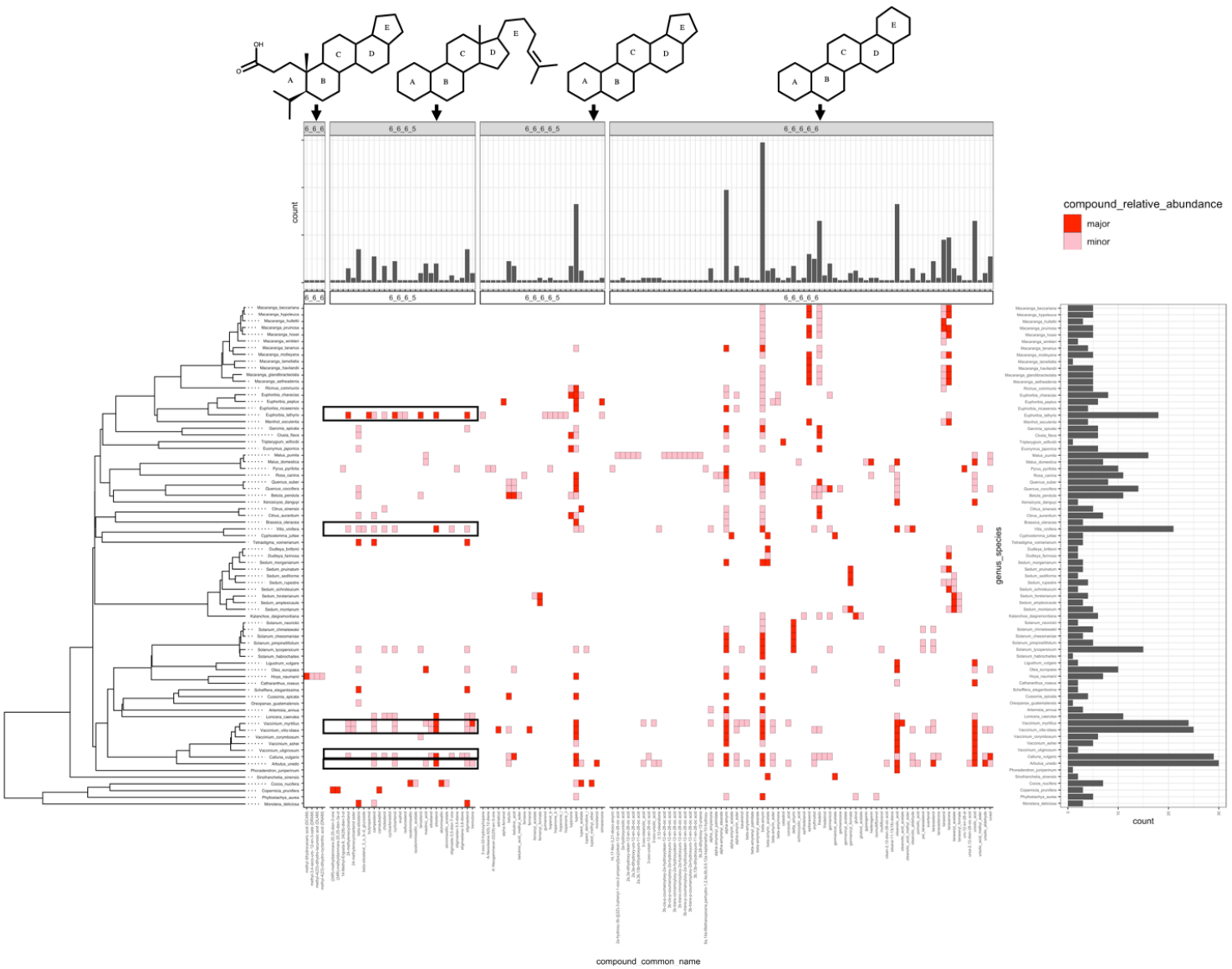
6_6_6
6_6_6_5
6_6_6_6_5
6_6_6_6_6
count
compound_relative_abundance
major
minor
genus_species
count
compound_common_name

Figure 8 (previous page). A comprehensive map of 696 triterpenoids occurrences across 76 different plant species derived from literature information. On the heat map in the middle, the presence of triterpenoid occurrences is categorized into two components: dark red for the major compounds and light red for the minor compounds. The x-axis shows all the 130 unique triterpenoid names grouped into four groups based on the carbon structure of their five (complete or incomplete) rings: 6-6-6-6-6, 6-6-6-6-5, 6-6-6-5-0 and 0-6-6-6-5; and the bar plot on top demonstrates the number of occurrences of each of these triterpenoids across all the studied plant species. The y-axis shows all the plant species grouped according to phylogenetic relationships (the relatedness of two adjacent species on y-axis is indicated by the length of the branch between them); and the bar plot on the right shows the number of documented occurrences of unique triterpenoids in each of these species.

3.1 Certain individual triterpenoids and triterpenoid groups are very common compared with the rest

3.1.1 Individual triterpenoids

Based on the data, it appears that some triterpenoids have been reported in a higher number of species than others; particularly like alpha-amyrin, beta-amyrin, oleanolic acid, ursolic acid, epitaraxerol, taraxerone, taraxerol, frieldelin, lupeol, sitosterol, beta-sitosterol and stigmasterol. These appear in a very wide range of genera/species compared to the rest of other reported (major or minor) triterpenoids (see Figure 7 and Figure 8). This could be because these triterpenoids are among the most ancient forms of plant triterpenoids. Since these triterpenoids all came from different group: alpha-amyrin, beta-amyrin, oleanolic acid, ursolic acid, epitaraxerol, taraxerone, taraxerol, frieldelin (6-6-6-6-6 group); lupeol (6-6-6-6-5 group); sitosterol, beta-sitosterol and stigmasterol (6-6-6-5-0 group), this may also mean that each group has ancient origins. Another possible reason for why these compounds being common could simply be due to detection efficiency, where these triterpenoids are more efficiently detected in analyzed samples by the instruments as compared to less common triterpenoids. This could lead to their prevalence on the maps. Another possibility is that these compounds are probably the most crucial among the triterpenoid class regarding their role in plant biological activities.

If only triterpenoids which are major are taken into account; then alpha-amyrin, beta-amyrin, oleanolic acid, ursolic acid, epi taraxerol, taraxerone, lupeol, sitosterol in addition to delta-amyrin are commonly observed across the most genera/species. A very likely reason for their abundance in quantity in many samples is that these triterpenoids have crucial chemical or biological roles in plants, whether it is for environmental adaptation, defense against harmful organisms or other functions. If the speculation about common triterpenoids being the primitive and original ones is true, it means triterpenoids like delta-amyrin could have come later in the evolutionary path but now have become very essential compounds in plant chemical activities.

Some triterpenoids like taraxerol, frieldelin, beta-sitosterol and stigmasterol are very common but they mostly appear as minor compounds (lower quantity within a plant body in comparison to major compounds). Taraxerol, frieldelin, beta-sitosterol and stigmasterol may be primitive or very detectable compounds in plants, but they are possibly traits from the evolutionary process, and they don't actually play a big role in plants' chemical function (not anymore at least). Another possible explanation for this data feature is that taraxerol, frieldelin, beta-sitosterol and stigmasterol do play a big role in plant chemistry, but mostly as transient intermediates or side-products- on biochemical pathways leading to triterpenoid compounds that are metabolic end points; so that's why they are only detected at a very small amount even though they are very common.

3.1.2 Groups of triterpenoids

The 130 unique triterpenoids across 41 genera /76 species are put into four groups based on the carbon structure of their five (complete or incomplete) rings: 6-6-6-6-6 group (74 compounds), 6-6-6-6-5 group (24 compounds), 6-6-6-5-0 (28 compounds) and 0-6-6-6-5 group (4 compounds) (see Figure 7 and Figure 8, top bar plot and Figure 9). From that, the 6-6-6-6-6 group is the most frequently reported across all documented compound groups which contain more than half of all documented unique triterpenoids (74 out of 130 compounds or 58%). Within this group, alpha-amyrin, beta-amyrin, and oleanolic acid followed by friedelin and ursolic acid are the most common compounds, and they clearly stand out from other compounds in the same group by the number of plant genera containing them. The 6-6-6-5-0 and 6-6-6-6-5 groups are next in line in diversity, each of them consisting of 28 and 24 unique triterpenoids, respectively. However, like the 6-6-6-6-6 groups, the 6-6-6-6-5 group has a few triterpenoids like lupeol followed by betulin and betulinic acid standing out as the most visible common compounds among more plant genera in comparison with other triterpenoids within the same group. Meanwhile, the 6-6-6-5-0 group does not have a few dominant unique compounds like the 6-6-6-6-6 group or the 6-6-6-6-5 group, but instead it has several triterpenoids including stigmasterol, beta-sitosterol, campesterol, cholesterol, cycloartenol and sitosterol … appearing in fewer genera equally. The 0-6-6-6-5 group only includes four triterpenoids methyl dihydrocanaric acid (DCAM), methyl-3,4-seco-urs-12-en-3-oate (DRAM), methyl-4(23)-dihydro-lacunosic acid (DLAM) and methyl-4(23)-dihydro-nyctanthic acid (DNAM); which are all documented within the species *Hoya naumanii*. In conclusion,

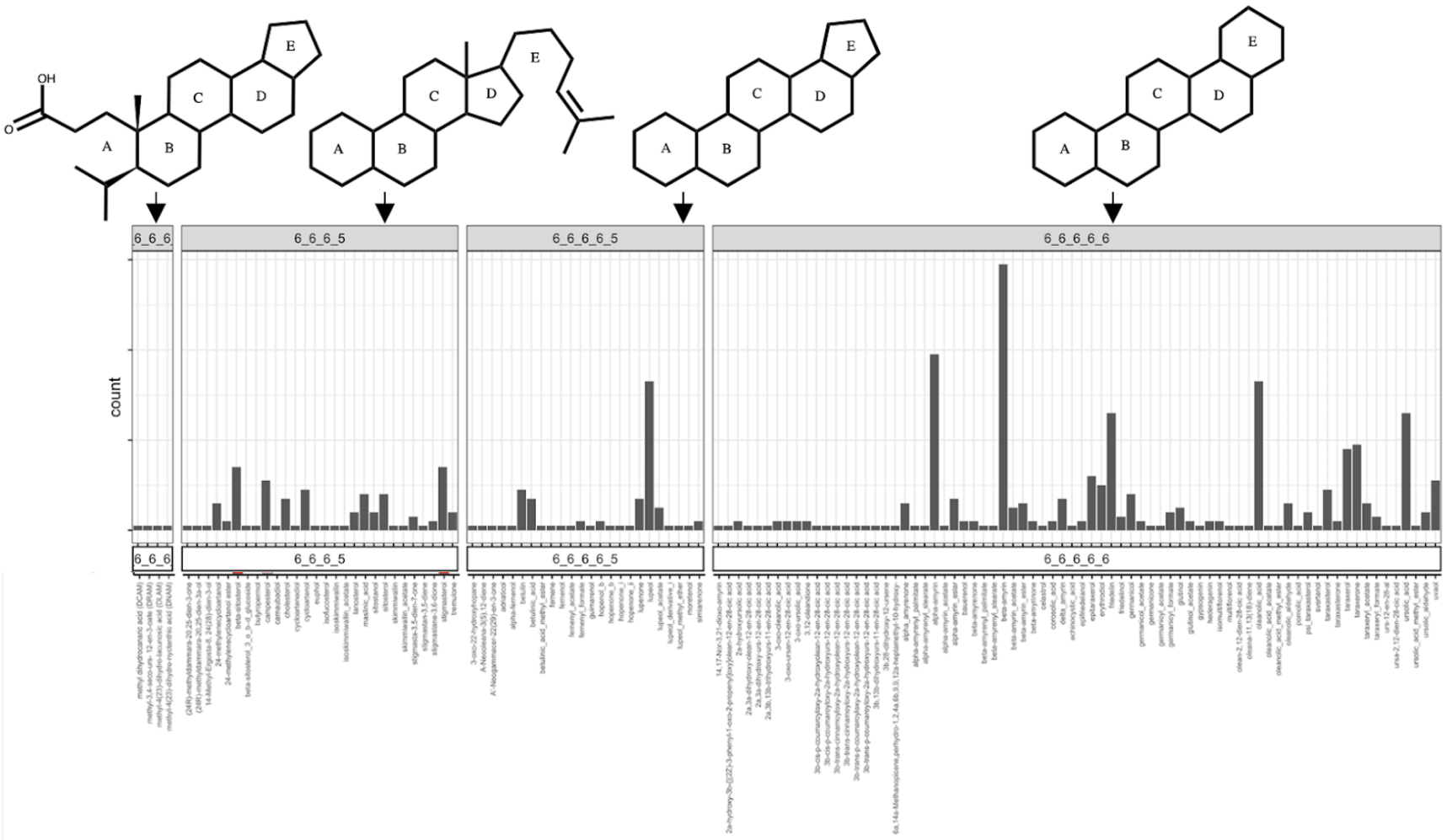

Figure 9. A compact figure of each triterpenoid by the number of their occurrences in each of the four groups from left to right: 0-6-6-6-5, 6-6-6-5-0, 6-6-6-6-5 and 6-6-6-6-6. Each tick on the y-axis indicates occurrences in five different species.

the great diversity of the 6-6-6-6-6 group (making up 57% of documented triterpenoids and including all the previously listed triterpenoids except lupeol) indicates the likely advantage of those molecular structures in plant function: the 6-6-6-6-6 might be more metabolically easy and efficient (requires less energy per molecule) to synthesize, or that the group can be converted into more useful molecules that gives plants a better fitness advantage to survive.

Another way of grouping all the triterpenoids besides their carbon ring structure is classifying them based on their main functional group, and the common functional groups being found in triterpenoids are alcohols, ketones and acetates. From RStudio analysis, it is found among the 696 triterpenoids occurrences: 415 are triterpenoid alcohols, 67 are triterpenoid ketones, 25 are triterpenoid acetate and 189 are triterpenoids with other less common functional group. From these given numbers, it can be said that statistically the ratio between triterpenoid alcohol: triterpenoid ketone: triterpenoid acetate is about 16.6:2.7:1. Triterpenoid alcohols therefore is definitely the most commonly observed (appear around 60 % of all triterpenoid occurrences) followed by triterpenoid ketone and triterpenoid acetate among all the documented triterpenoid occurrences. This ratio is a significant piece of information that will be used later for a comparison in Chapter 3.

3.2 Relative to distantly related genera/species, more closely related genera/species have more similar triterpenoid profiles

3.2.1 Genus level

In the genus-level map of triterpenoid occurrence (see Figure 7), the closely related genera *Vaccinium*, *Calluna*, *Arbutus* and *Lonicera* all produce eight similar triterpenoids and (18 similar triterpenoids if *Lonicera* is excluded). The genera *Vaccinium*, *Calluna* and *Arbutus*, *Lonicera* and *Artemisia* (on the same branch with *Lonicera*) all have alpha-amyrin and beta amyrin as major compounds, except for *Lonicera* which has them as minor compounds. Besides that, *Vaccinium*, *Calluna Arbutus* and *Lonicera* also have sitosterol, lupeol and ursolic acid as major compounds (except for *Lonicera* which has no lupeol and has ursolic acid as a minor compounds) and have campesterol, cycloartenol, and stigmasterol as minor compounds. These four genera all have oleanolic acid, with *Vaccinium* and *Arbutus* have it as a major compound and *Lonicera* and *Calluna* have it as a minor one. In addition to that, all six genera on the top branch of the map including *Clusia*, *Garcinia*, *Macaranga*, *Ricinus*, *Manihot* and *Euphorbia* all share beta-amyrin (a major compound for *Macaranga* and *Garcinia* and a minor compound for the rest). With the exception of *Manihot*, the other five genera also contain alpha-amyrin (a major compound for *Macaranga* and *Garcinia*) and lupeol (a major compound for *Ricinus* and *Euphorbia*). Another observed trend on a smaller scale is with the two genera *Betula* and *Quercus*, which share up to eight triterpenoids: they both have beta-sitosterol, beta-amyrin, erythrodiol, friedelin and oleanolic acid; *Betula* have betulin and betulinic acid as major compounds and lupeol as a minor compound and vice versa for *Quercus*.

As a result, more closely related genera tend to have similar triterpenoid composition when it is compared to those that are further apart. Moreover, this trait is even more visible when the genera containing a lot of triterpenoids are being compared, like in the case of *Vaccinium*, *Calluna* and *Arbutus* (each has at least 29 unique triterpenoids). These are very likely traits being observed as these genera develop over their evolutionary path, although small divergences can start to be seen. For example, of the 18 triterpenoids that *Vaccinium*, *Calluna* and *Arbutus* share, oleanolic acid and taraxasterol are both major compounds in *Arbutus* and are both minor compounds in *Calluna*; and oleanolic acid is a major compound but taraxasterol is a minor compound in *Vaccinium*. A possible explanation is that the three genera share a recent common ancestor; but as evolution occurs each genus needs to adapt to its environmental condition and they start to function in different ways: both oleanolic acid and taraxasterol are still crucial for *Arbutus* activities so they are retained; *Vaccinium* activities only emphasized olenolic acid for it biological activities so taraxasterol is phasing out; *Calluna* is losing both mentioned triterpenoids as it can likely function with minimal presence of the two compounds.

3.2.2 Species level

In Figure 8 and Figure 10, four out of five documented *Vaccinium* species (*V. myrtillus*, *V. vitis-idaea*, *V. corymbosum*, *V. ashei* and *V. uliginosum*) and the related *Calluna vulgaris* and *Arbutus unedo* produce large amounts of alpha-amyrin, beta-amyrin, oleanolic acid and ursolic acid along lupeol as a major compound; except for *V. ashei* has lupeol and beta-amyrin as minor compounds and *V. uliginosum* which lack alpha-amyrin,

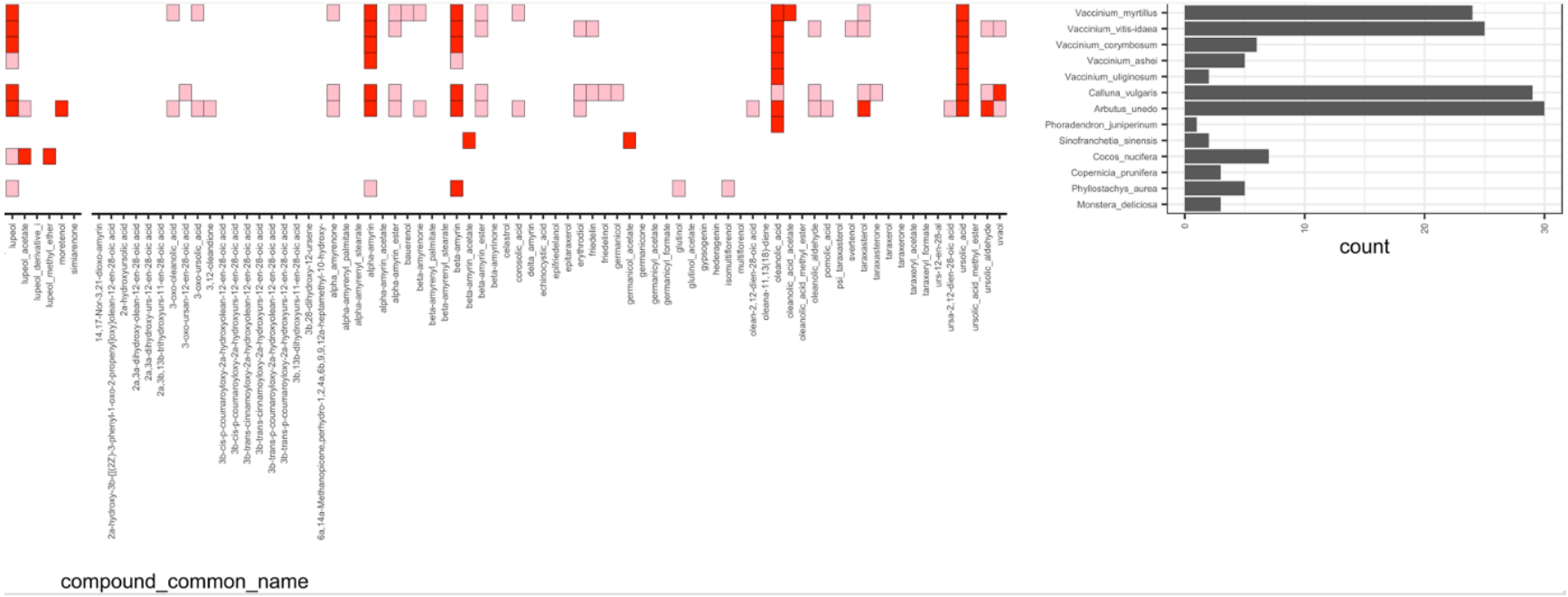

Figure 10. A visible trait reflecting divergent evolution is observed among triterpenoids in the *Vaccinium* species and the related *Calluna vulgaris* and *Arbutus unedo*.

beta-amyrin and lupeol. Also, the *Solanum* species (*S. neorickii*, *S. chmielewskii*, *S. pimpinellifolium*, *S. habrochaites* and *S. lycopersicum*) similarly have presence of alpha-amyrin and beta-amyrin except *S. neorickii* and *S. habrochaites* which do not have alpha-amyrin. And except *S. habrochaites*, all of them also have delta-amyrin as a major compound. Another different trend is observed in the *Macaranga* species (*M. beccariana*, *M. hypoleuca … M. aetheadenia*) and the related *Euphorbia* species (*E. characias*, *E. peplus*, *E. nicaeensis* and *E. lathyris*) and *Ricinus communis*: they all share beta-amyrin as a minor compound except *M. tanarius* which has beta-amyrin as a major compound. The majority of *Macaranga* species also have epitaraxerol (9 out of 12 species) and taraxerone (8 out of 12 species) as major compounds, and friedelin and taraxerol (10 out of 12 species) as minor compounds, though *Euphorbia* species and *Ricinus communis* don't show these traits. In contrast, they (*Euphorbia* species and *Ricinus communis*) all share lupeol as a major compound, while the *Macaranga* species lack this. As these trends are very similar to those being observed on the genus map, it is likely that closely related species also have very similar triterpenoid composition. Besides that, the species within a genus usually share many similar triterpenoids with each other compared with the species outside their genus, and this is clearly demonstrated by the species in the genera *Macaranga*, *Euphorbia*, *Solanum* and *Vaccinium*. The explanation for this, through an evolutionary perspective, is probably the same for that of the genus map: all the species within a genus and sometime very close species on the branch (like the *Vaccinium* species and *C. vulgaris* and *A. unedo*) perhaps share a common recent ancestor, and as each species tries to adapt to their unique environment their biological function also starts to branch. Small differences in the triterpenoid composition of their

species arise even though overall, their triterpenoid profiles are still overwhelmingly similar to each other.

3.3 More closely related genera/species do not necessarily have similar diversity levels

3.3.1 Genus level

Taking an overall look into the bar plot on the right of Figure 7, the three related genera *Vaccinium*, *Calluna* and *Arbutus* all contain more than 25 different triterpenoids compounds, indicating very high diversity of triterpenoids. Another similar case is that of the five genera *Pyrus*, *Malus*, *Rosa*, *Betula* and *Quercus* on two close branches all contain at least ten different triterpenoids each. In contrast, the genera *Euphorbia* and *Manihot* are also very closely related; but while *Euphorbia* has an extremely diverse triterpenoid profile of up to 28 different compounds, *Manihot* only has four different triterpenoids. In a similar fashion, the genus *Vitis* shares a branch with the genus *Cyphostemma*; but *Vitis* has more than 20 different triterpenoids while *Cyphostemma* only has three. As a result, correlation between triterpenoid diversity and phylogenetic relationship of plant genera is not always true. Although it is true from studying trends earlier that genera which are closer on trees are more likely to have similar triterpenoid composition, these genera's diversity of triterpenoids are not always the same: two close genera can share several same triterpenoids, but one might have a lot of extra triterpenoid compounds while the other just have those similar ones.

3.3.2 Species level

When the bar plot is observed on the species level in Figure 8, some correlation between triterpenoid diversity level and phylogenetic traits can also be seen, similar to that of the map on the genus level. All the *Vaccinium* species and the related *Calluna vulgaris* and *Arbutus unedo* contain a significant number of different triterpenoids, with all species except *V. uliginosum* having at least five compounds (see Figure 10). Similarly, all of the 8 Sedum species (*S. morganianum*, *S. pruinatum*, *S. sediforme*, *S. rupestre*, *S. ochroleucum*, *S. forsterianum*, *S. amplexicaule*, *S. montanum*) and all the *Macaranga* species show a low level of triterpenoids diversity with none of them have five more than five compounds. These mentioned examples show a strong correlation between triterpenoids diversity and the phylogenetic relationship of the plant species. On the other hand, similar levels of diversity were clearly not being observed between some species within a genus. Looking at the triterpenoid diversity across all the *Solanum* species, the species *S. lycopersicum* has 15 different unique triterpenoids even though all other species in the group contain no more than five compounds. Another case is within the *Euphorbia* genus: *E. lathyris* contains 18 unique triterpenoids even all other *Euphorbia* species contain no more than eight compounds. These cases help to confirm that the correlation between triterpenoid diversity and phylogenetic relationship is not always the case between species within the same genus or from closely related genera. Although it has been argued earlier that the closely related species seem to have more similar triterpenoids composition, the unbalanced diversity between species within even the same genus is not uncommon. A species like *E. lathyris* can have many additional triterpenoids

beside the similar triterpenoid that it shares with other species in the same genus; or the opposite can happen where a species like *V. uliginosus* only have olenolic acid and ursolic acid (which all the documented *Vaccinium* species have) but it lacks most other triterpenoids which all other species in the same genus possess.

3.4 Data feature of convergent traits that was observed on the species level

Based on section 3.2, there is a strong correlation in triterpenoids composition between closely related genera/species in most cases on the maps; and this demonstrates the common divergent evolution path of the documented plants. However, sometimes distantly related species have similar chemistry through more convergent traits: a species shares several triterpenoids with a very distantly related species on the phylogenetic trees, although none of the other species in its genus possesses these compounds. These traits can be observed mostly on the species level map since a genus will be shown to have a triterpenoid on the genus level map even if only one species within the genus contains that triterpenoid, so convergent evolution traits are harder to observe and detect.

The two species *E. lathyris* and *V. vinifera* are very distantly related and also don't really share any common triterpenoids in the 0-6-6-6-5 group, the 6-6-6-6-5 group and the most dominant 6-6-6-6-6 group. However, they both share up to six similar triterpenoids of the 6-6-6-5-0 group (see Figure 11). Besides that, E. lathyris produces many triterpenoids that are the same as those produced by a group of its distant relatives. These triterpenoids include: five similar triterpenoids with *V. myrtillus* and *V. vitis-idaea*, six similar

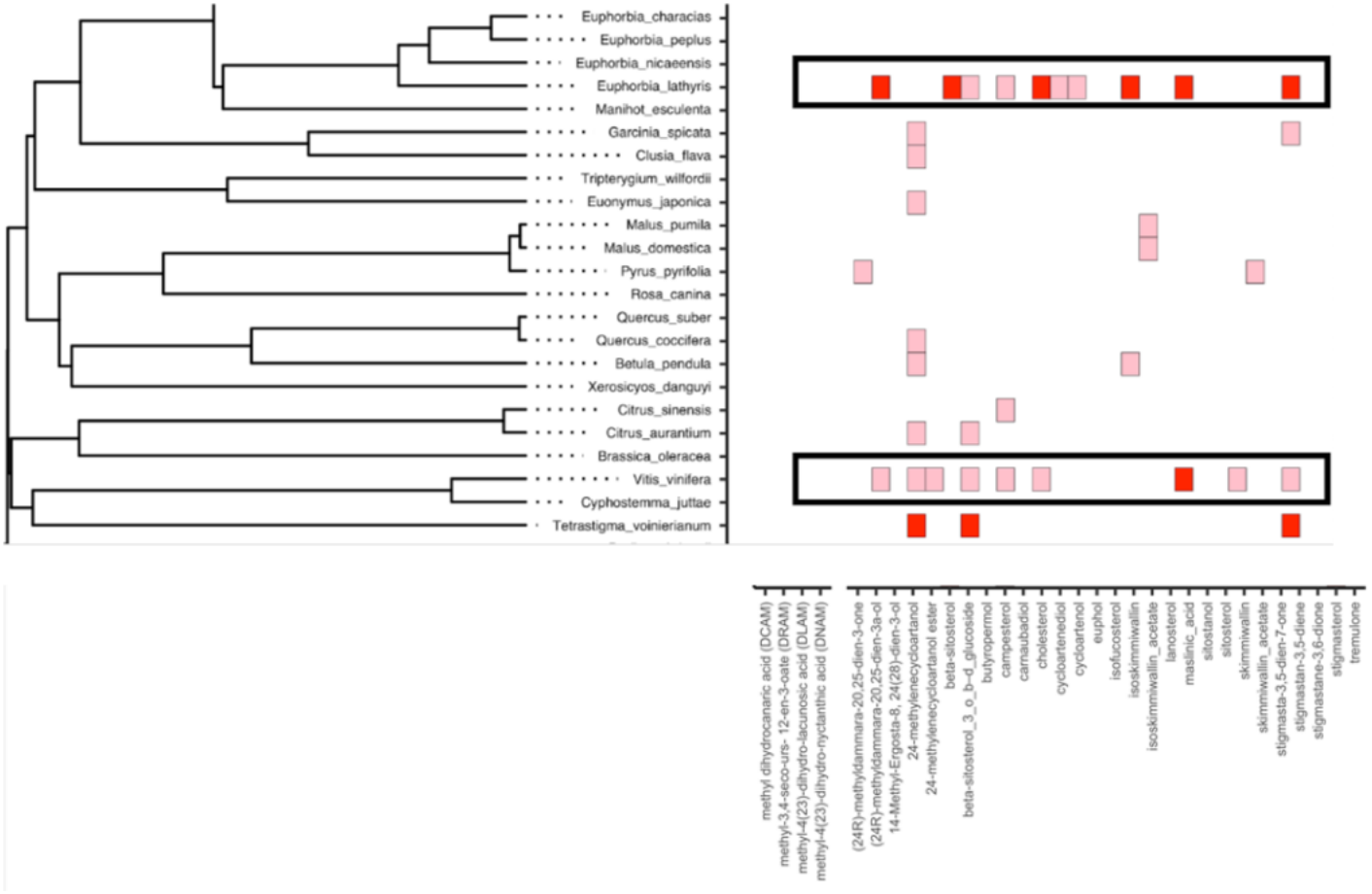

Figure 11. A visible trait reflecting convergent evolution is observed among triterpenoids between the two species *E. lathyris* and *V. vinifera*.

triterpenoids with *C. vulgaris* and four similar triterpenoids with *A. unedo*. On a smaller scale, *E. lathyris* also have four similar triterpenoids with the species *S. lycopersicum*. For *V. vinifera*, the species share a number of similar triterpenoids with the previously mentioned group of species which is not closely related to *V. vinifera*: five similar triterpenoids with *V. myrtillus* and *V. vitis-idaea*, seven similar triterpenoids with *C. vulgaris* and four similar triterpenoids with *A. unedo*. *V. vinifera* also have four similar triterpenoids with the species *S. lycopersicum*. Finally, *E. lathyris*, *V. vinifera* and other species described above all share triterpenoids in the 6-6-6-5-0 group. Their triterpenoid profiles are very different in the 6-6-6-6-5 group and the most dominant 6-6-6-6-6 group, where most related species share similar triterpenoids. Consequently, though more closely related species often have more similar triterpenoid chemistry compared to more distantly related species through divergent evolution; sometimes very distantly related species (like E. lathyris and V. vinifera) can also have remarkably similar triterpenoid profiles. This could be explained by convergent evolution. In these cases, the plant species, instead of reducing production of some triterpenoids inherited from their ancestors, start to produce certain triterpenoids which they require to function in their environment. That means two non-related plant species can have some similar triterpenoids on some occasions because they live in a similar environment that compels their biochemical activities to function similarly. This explanation seems to be even more likely considering that the example species described above all share triterpenoids in the 6-6-6-5-0 group, and triterpenoids profiles in their 6-6-6-6-5 group and the most dominant 6-6-6-6-6 group are quite different.

4. Conclusion

In the end, the major findings from this chapter can be summed up in three major points: (i) some triterpenoids are much more common than the other, (ii) some triterpenoids appear to have undergone divergent evolution, and (iii) some triterpenoids appear to have undergone convergent evolution. In the first point, it is very interesting to see how prevalent certain triterpenoids (either with major or minor amounts) are among the chemical composition of the documented plants: alpha-amyrin, beta-amyrin, oleanolic acid, ursolic acid, epitaraxerol, taraxerone, taraxerol, frieldelin (6-6-6-6-6 group); lupeol (6-6-6-6-5 group); sitosterol, beta-sitosterol and stigmasterol (6-6-6-5-0 group); as it seems likely that they are very primitive or detectable compounds. These triterpenoids, with the exception of taraxerol, frieldelin and beta-sitosterol, can be found in significant amounts in most plant species, indicating they are still important biomolecules for diverse plant species. In addition, the common presence of taraxerol, frieldelin and beta-sitosterol as minor compounds is also worth studying and being understood. A possible reason for this phenomenon is that taraxerol, frieldelin and beta-sitosterol might be compounds which were once useful for the ancestors but not so much for the modern plants' biochemistry; but it seems more likely that these are important but intermediate compounds which are only present for a short time in the reaction chains in plant chemistry. Among the triterpenoid groups, it is also very interesting to see how the 6-6-6-6-6 group is so dominant in diversity (more than half of all documented triterpenoids) and also contains most of the abundant triterpenoids (alpha-amyrin, beta-amyrin, oleanolic acid, ursolic acid, epitaraxerol, taraxerone, taraxerol, frieldelin). This strongly indicates

that the general structure of the 6-6-6-6-6 group is very heavily favored by plants due to a number of possible reasons: they are easy to synthesize, they are crucial for biochemical functions, they can be precursors that can be converted into useful molecules for those functions.

From both the genus and species level maps, it does seem true that genera/species which are closer to each other are more likely to have similar triterpenoids than those that are further apart, and this is even more visible when genera/species containing lots of triterpenoids are observed. However, it is important to know this trend is not true when it comes to triterpenoid diversity: close genera/species on the phylogenetic trees may share many similar triterpenoids, but the number of total triterpenoids they possess can be drastically different. This is also true in terms of each triterpenoid's quantity: it is common that the many triterpenoids which are shared between the genera come at different amounts-minor for one genus and major for another. The reasons for these particular trends can be explained with plant divergent evolution: many genera/species that develop from one common ancestor have the same triterpenoids, but over time certain triterpenoids slowly start to be produced in a smaller amount or completely disappear to adjust to the need to function effectively in each of their environments or niches when they branch out, even though overall they still share lots of unique triterpenoids which make their triterpenoids overwhelmingly similar.

On the other hand, the species map (see Figure 8) stands out from the genus map (see Figure 7) in a way that it allows for the observation of the rarer convergent trait in plant

triterpenoids. Although this is a lot less common and was only seen among six species out of all the documented plants in this project, the situation happens when very distant related species appear to share many more similar triterpenoids than usually observed. This can be explained very well through the convergent evolution lens, in contrast to the common divergent traits: instead of reducing and losing triterpenoids, some similar triterpenoids are being produced in greater amount by the non-related species because they need those certain triterpenoids to adapt to their similar environments. On the side notes, it is also interesting to see convergent traits appearing exclusively with the 6-6-6-5-0 group, and the divergent traits seems to be more commonly demonstrated with the triterpenoids from the 6-6-6-6-5 group and the most dominant 6-6-6-6-6 group.

From the observed traits and the speculations that pointed to divergent and convergent evolution of plants regarding their triterpenoid profile, a question can be raised about the relationship between triterpenoids produced by plants and their environmental equation. As being mentioned, plants with a certain initial triterpenoid composition can evolve in two ways: producing some triterpenoids in a smaller amount or not producing some triterpenoids altogether (through divergent path); or starting to produce new triterpenoids it never has (through convergent path). However, it is important to point out that the common possible reason for both scenarios to happen is because plants adjust their triterpenoid composition in order to adapt and thrive in the condition in which they are growing. This alternate hypothesis about association between triterpenoids and plant environmental conditions is still largely untested with studies and documentation having

only been done in some specific and isolated cases, so the information obtained from this chapter provides a great opportunity for future experiments to be conducted.

Chapter 3: Experimental Analysis

1. Introduction

Triterpenoids are not the only compounds that can be found on plant surfaces. Linear aliphatic compounds including alkanes, fatty acids, fatty alcohols, aldehydes, and ketones can also be found. In some cases, waxes accumulate to such high abundances that they are visible to the naked eye as a white, powdery coating. These highly abundant coatings are called “wax blooms” (Burow et al. 2009). The goal of this chapter is to assess triterpenoid occurrence (as well as the occurrence of other compounds) in wax blooms. It is anticipated that this dataset will be, to some degree, comparable with the literature dataset from the previous chapter. This also raises two main questions. The first question is whether a newly collected dataset from wax blooms will produce the same arguments made from the literature meta-analysis in Chapter 2. The issue is that all the literature sources focused heavily on triterpenoids in plants, and it is a great thing since this extensive information then can be better used in combination to produce the comprehensive triterpenoid data set. On the other hand, focusing heavily on triterpenoids means that these sources can be biased toward just plants with triterpenoids, so it is not known if triterpenoid presence actually occurs that often within surface waxes of all plant species. The second question is, in a newly collected dataset from wax blooms, what other compounds can also be found along with triterpenoids in waxes and if these compounds interact or have any influences on the triterpenoids being produced. This second question also cannot be answered with the previous literature chapter that targets

heavily on triterpenoids. Because of these two unanswered questions, there is an opportunity to collect information from a set of actual plant samples and analyze information the same way triterpenoid documentation was done in the previous chapter. If this is accomplished, it can help to support the arguments made in Chapter 2 about some triterpenoids being very common compared to others and some triterpenoids seeming to have undergone convergent or divergent evolution. In addition, this wax bloom dataset can help to improve knowledge of other compounds in plant surface waxes alongside triterpenoids, and in the long term, it can also contribute to building ideas for future experiments to understand plant biochemistry and to realize triterpenoids' potential in activities like agriculture.

This chapter will be an experimental analysis chapter, in contrast to the literature review Chapter 2. The information used for building this data set will come from various plant samples obtained through citizen-scientist acquisition method; and all the analyzed species are collected in a random fashion so there will not be any biases toward just plants containing triterpenoids. The objective is to build a map of surface wax compound occurrences in various plant species on a large and sizable scale just like the maps presented in Chapter 2, with the difference being that experimental information will be the source used to build this map.

The compound occurrence map described in this chapter will allow for the analysis of triterpenoids with other compounds in surface waxes from a newly obtained set of 130 experimental samples. As a result, good analytical observation in many different ways

can be made with trends in a broader and more general map including triterpenoids and also other compounds present in surface waxes, and potentially trends of their interaction in this map as well.

2. Methods

The goal in this chapter is to build a map of triterpenoid occurrence based on at least 100 plant samples that were analyzed with gas chromatography-mass spectrometry (GC-MS). This map is intended to be more quantitative than the maps built in Chapter 1 with literature data and will be constructed via three major steps: (i) sample preparation, (ii) GC-MS analysis, (iii) data analysis and construction of a plant chemistry map by RStudio.

2.1 Sample documentation & preparation

Many samples of different plant species across the US were sent to Dr. Busta's lab by citizen scientist collaborators. These samples were collected on cotton swabs, and it was necessary to document all of them first. A spreadsheet was created for this purpose and each sample that arrived was given an ID number. The ID number of each sample was recorded along with the collector's name, the state of origin, and most importantly, the plant species' scientific and common name. After the first documenting step is finished, the samples were extracted and prepared for GC-MS analysis. Each GC vial was labeled according to the ID of its sample and 1 mL of chloroform was added to each vial. Each

swab was removed from its pouch and was immersed in the respective vial containing chloroform for 1 minute. After that, the swabs could be discarded, and all the chloroform could evaporate. When chloroform was confirmed to be completely evaporated, 125 uL of pyridine and 125 uL of N,O-Bis(trimethylsilyl)trifluoroacetamide (BSTFA) were added to the vials to convert triterpenoid alcohols into trimethylsilyl derivatives. Doing this helped to increase volatility and enabled MS fragmentation of triterpenoid molecules in a way that it was easier to identify their structures. Then they were capped, vortexed, and finally incubated at 70°C for 45 minutes; and after that they were ready for GC-MS analysis.

2.2 Acquiring the GC data

Prepared GC samples were run on the HCAMS 109 GC-MS system using a GC method developed for this project. GC-MS analysis was done using standard 70 eV EI ionization with chromatographic conditions as follow: 50°C hold for 2 minutes, 40°C /minute ramp to 200°C, hold for 2 minutes, 3 °C/minute to 320°C, hold for 2 minutes. Since the concentration of analytes in the sample was not known (they were collected by citizen scientists), an initial injection volume of 1 uL was utilized. After a sample was run, the resulting total ion chromatogram (TIC) was inspected and used to categorize the sample TICs as “good”, “bad”, or “ok” (see Figure 12). The “good" ones were TICs with very clear and much larger (higher than end baseline) target peaks; and the “ok” were TICs with some peaks of contaminants and target peaks were often not higher than the end of run baseline. The “bad” ones were samples with TICs where peaks were essentially

absent and only noise could be observed so improvement was unlikely. In general, all the "ok" and some of the "good" ones were most likely to be put on a rerun for improvement, with adjustment of the injection volume. If the peaks stood out from the baseline but were not visible enough (lots of noise along the baseline) or stood lower than the "ramp baseline" (can be seen within 45-50 minutes retention time parts of the TIC in Figure 12), the injection volume was increased. In addition, peaks were expected to fall within the intensity range of around 10^6, so injection could also be increased or decreased if the peaks went out of this range. At the same time, it should be noted that the injection volume was kept to not exceed 3 uL since this could overload the instrument's injection port liner. The purpose of these reruns was to maximize the number of GC samples with "good" data quality.

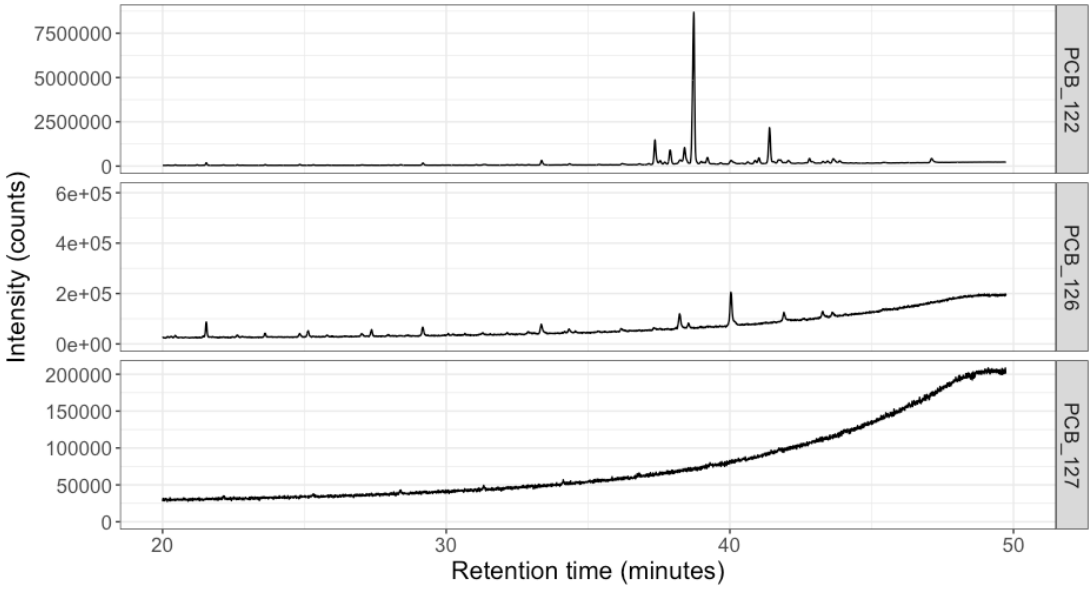

Figure 12. An example of how GC sample TICs are being categorized: a "good" TIC (PCB_122 on top), an "OK" TIC (PCB_126 in the middle), and a "bad" TIC (PCB_127 at the bottom).

2.3 Analyzing the GC data, building a spreadsheet of wax bloom information and constructing a plant chemistry map by RStudio

Efforts have been made to run and rerun to collect data from as many "good" PCB samples as possible. After that, the TICs of these "good" samples were exported as AIA files into a flash drive. With the software RStudio along with the code posted in **Appendix B**, the next step was to build a comprehensive plant chemistry map. First, RStudio helped to select and integrate all the principal and crucial peaks from the TICs in the drive, and to observe their respective mass spectra. With a mass spectra library of common organic compound groups in plants provided by Dr. Busta, these peaks could be identified and thus the main chemical components (alkanes, alcohols, aldehyde, fatty acids, triterpenoids …) along with their relative abundances percentages within each sample could be documented. After that, RStudio allowed for these integration results to be written into multiple CSV files (one CSV file per PCB sample), and then combined these integration CSV files with the ID sample spreadsheet in step 2.1 into a single Google Excel sheet. With the final summarized Google Excel sheet, it was possible to build a chemical composition map of various plant species. All the species were organized along their phylogenetic relationship while all their main chemical components were put into two groups: triterpenoids and other compounds (alkanes, alcohols, aldehyde, fatty acids). As this map was constructed, interpretation about trends of triterpenoids in plants could be observed and then discussed.

In addition, GC-MS and gas chromatography-thermal conductivity detector (GC-TCD) measurements were performed on three different triterpenoid groups: ketone group (represented by beta-amyrone), alcohol group (represented by beta-amyrin) and acetate group (represented by beta-amyrin acetate). Utilizing the available PCB samples containing triterpenoids, four different mixture samples were prepared with each containing all three mentioned triterpenoids and then analyzed by GC-TCD and GC-MS. The triterpenoid molar abundances in these samples were not quantified as the aim is making a comparison of each triterpenoid's abundance ratio between the two instrumental systems. The integrated areas under identified chromatogram peaks for these triterpenoids were consequently collected and then a correction factor was calculated, so comparisons can be made between GC-MS and GC-TCD data for the three different triterpenoids.

3. Results & Discussion

In order to observe and understand surface wax chemistry in collected samples from many plant species, a thorough chemical analysis process was conducted. This process included two major steps: (i) preparation & documentation of collected samples, GC-MS data acquisition, chemical map construction (section 3.1), and (ii) results interpretation (section 3.2).

3.1 Data and results generation

A total 228 samples of 140 plant species were sent by citizen scientist collaborators to Dr. Busta's lab. They were all documented by ID numbers from PCB_1 to PCB_228 in a detailed Google Excel spreadsheet “wax_bloom_chem_data”. Then they were carefully prepared and analyzed with the GC-MS instrument. After all the samples were run through GC-MS and many attempts were made to rerun the “ok" TICs, some “ok” TICs were improved enough that they could be moved to the “good" category while the rest of the TICs were moved to the “bad" category. This resulted in 98 “bad" TICs from 61 species and 130 “good” TICs from 85 species, which these “good" ones can then be used for further analysis. The next step of integrating peaks using RStudio from the retained 130 “good” TICs produced 636 different selected peaks, which correspond to the occurrences of chemical compounds in these samples. Among these peaks, observation of their mass spectra with the help of the mass spectra library allowed for the identification of 595 peaks (or 93.6 % of total integrated peaks). Besides that, along with the identification of these peaks of compound occurrences was the documentation of their relative abundance (in percent) within each sample. All of this information was written into multiple CSV files (one CSV file per PCB sample) and then combined with “wax_bloom_chem_data” into a single Google spreadsheet.

Like the literature information obtained in Chapter 2, contradicting information about chemical composition from different samples of the same species was also recorded. This happened in collected samples from across 19 different plant species. As these situations were carefully inspected, three scenarios were discovered, and three respective solutions

were done to solve them. In seven species, there were slight differences in chemical composition between the PCB samples where more than 50% of the compounds in each sample are the same, so the solution is to take an average on the relative abundance values between samples in each species. A second scenario happening in nine species is where there are significant differences in the chemical composition between the PCB samples (fewer than 50% of compounds in each sample are the same); but careful investigation revealed that several samples were containing contaminants so these contaminated PCB samples can be filtered out and the rest was kept to represent these species. A third scenario occurring in three species is where there are significant differences in composition between samples from the same species but with no noticeable contaminants; so notes were taken of them for future studies (by using more samples as tiebreakers eventually) and they were filtered out of the dataset. Furthermore, ten species were dropped because their information within the phylogenetic tree library is not available to be included into the map. Ultimately, 376 documentations of compound occurrences including 16 unique triterpenoids and 64 other unique compounds across 72 plant species were retained.

With this final dataset, RStudio was used to build a map of chemical compound occurrences in collected plant samples (see Figure 13). This map shares many similarities with the maps created in Chapter 2 with four main components: a phylogenetic tree of all reported 72 plant species on the left; a heat map indicating compound abundance in each respective species; a bar plot on top showing the number of species in which each compound has been found; and a bar plot on the right showing the number of compounds

each species has been reported to synthesize. On the other hand, there are two noticeable differences between this map and the two literature data maps in Chapter 2. First, this heat map section was a lot more quantitative with the compounds not defined simply as "major" or "minor" compounds, but they came with relative abundance within each sample through the color intensity. In addition, the map also demonstrated a wide range of chemical compounds and not only the triterpenoid groups: triterpenoids and six compound groups (alcohol, aldehyde, alkane, fatty acid, ketone, and other compounds) were separated along the x-axis into groups, represented by six different colors on the bar plot on the right. At this point, trending relationships can be analyzed in two different ways: the overall trend of the main components (heat map of all the species/genus + phylogenetic tree) and the top plot, the trend between the main components and the right plot.

It is also important to talk about a potential pitfall about instrumental efficiency: GC-MS instruments being used might be more efficient in detecting some triterpenoids over the others. Consequently, an experiment was conducted to test this possible problem by using a GC-TCD instrument and a GC-MS instrument to analyze four mixture samples with each containing beta-amyrone, beta-amyrin and beta-amyrin acetate; and then the data of three triterpenoids were compared. The results of this mini-experiment regarding instrumental efficiency can be seen from Table 1: GC-MS is slightly more efficient at detecting ketone and acetate triterpenoids, while GC-TCD is slightly better at detecting alcohol triterpenoids. With these ratios being known at this point, it is possible to take them into account when specific data trends are discussed later.

Table 1. Average normalized peak area ratio for triterpenoids from three different groups being measured by GC-MS and GC-TCD.

	GC-MS	GC-TCD (20:1 split and 2 ul injection)	GC-TCD (5:1 split and 3 ul injection)
Beta-amyrone	4.87 (487%)	3.188 (318%)	3.425 (342%)
Beta-amyrin	1	1	1
Beta-amyrin acetate	1.855 (185%)	0.853 (85%)	0.898 (89%)

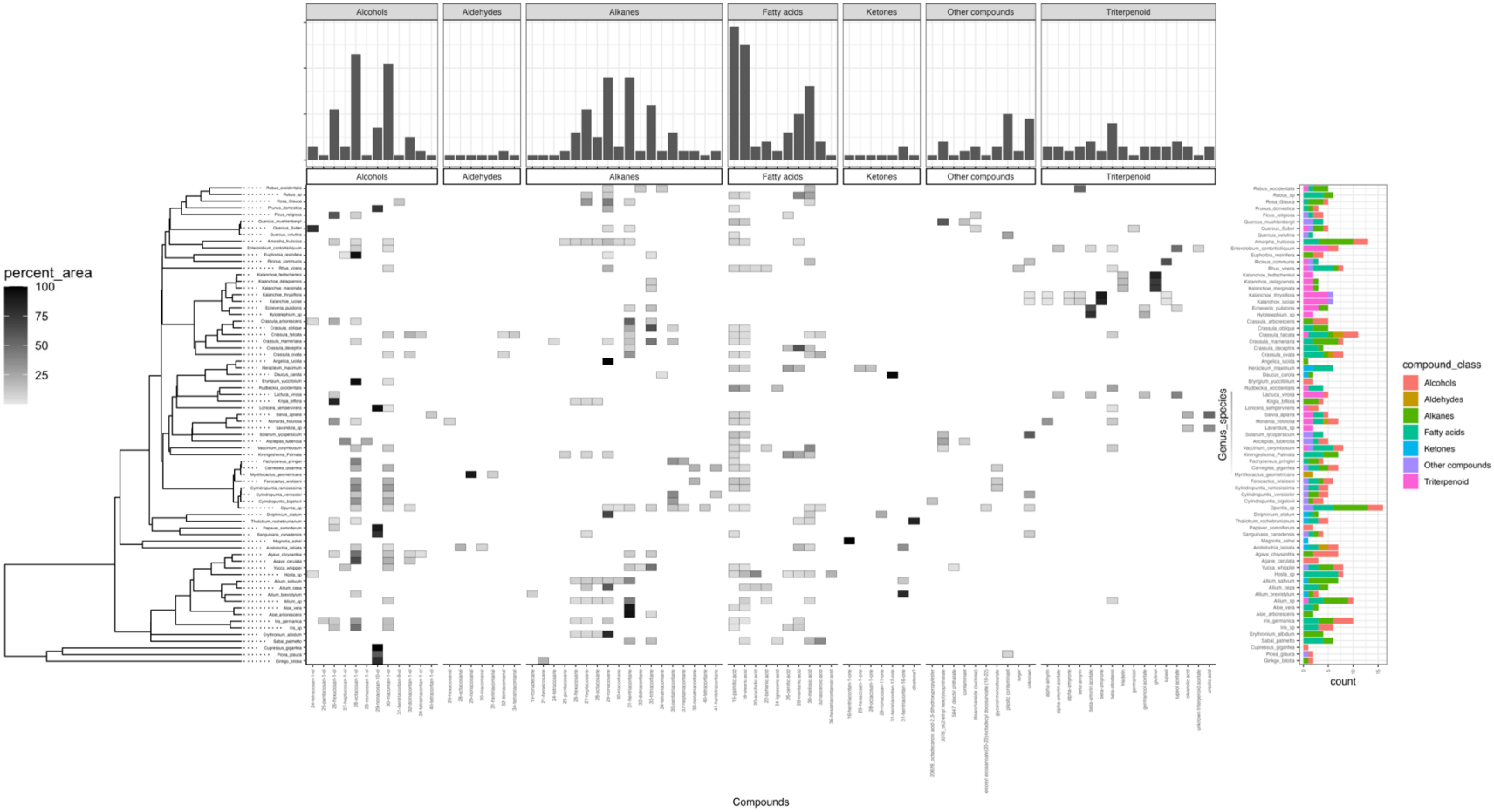
Alcohols
Aldehydes
Alkanes
Fatty acids
Ketones
Other compounds
Triterpenoid
percent_area
100
75
50
25
Genus_species
compound_class
Alcohols
Aldehydes
Alkanes
Fatty acids
Ketones
Other compounds
Triterpenoid
count
Compounds

Figure 13 (previous page). A comprehensive map of 367 documentations of plant wax bloom compounds from samples collected by citizen-based scientists across 72 different plant species. On the heat map in the middle, the presence of compound occurrences is shown quantitatively through the color intensity: darker shades indicate larger percentages of the compounds in quantity within each sample. The x-axis shows all the 64 unique compounds grouped into seven groups: alcohols, aldehydes, alkanes, fatty acids, triterpenoids and other compounds; and the bar plot on top demonstrates the number of occurrences of each of these compounds across all the studied plant species. The y-axis shows all the plant species grouped according to phylogenetic relationships (the relatedness of two adjacent species on y-axis is indicated by the length of the branch between them); and the bar plot on the right shows the number of documented occurrences of unique compounds in each of these species. In this right bar plot, the seven different colors correspond to the seven respective compound groups in each species.

3.2 Results interpretation

3.2.1 Certain individual compounds appear in a wide range of species

It seems that 16-palmitic acid and 18-stearic acid are found within most plant species across the phylogenetic trees and are not limited to any group or genus of plant species. Similarly, 28-octacosanol and 30-triacontanol in the alcohol group are the most common compounds after 16-palmitic acid and 18-stearic acid (Haslam et al. 2013) and they can also be found across most plant species (see Figure 14). A possible reason is likely that 16-palmitic acid, 18-stearic acid, 30-melissic acid, 28-octacosanol, and 30-triacontanol are the most metabolically efficient to synthesize and are also useful in biochemical function. This is briefly mentioned in literature where compounds 16-palmitic acid and 18-stearic acid are found to play major parts in the synthesis of phospholipids and glycolipids that make up the bulk of cellular membrane.

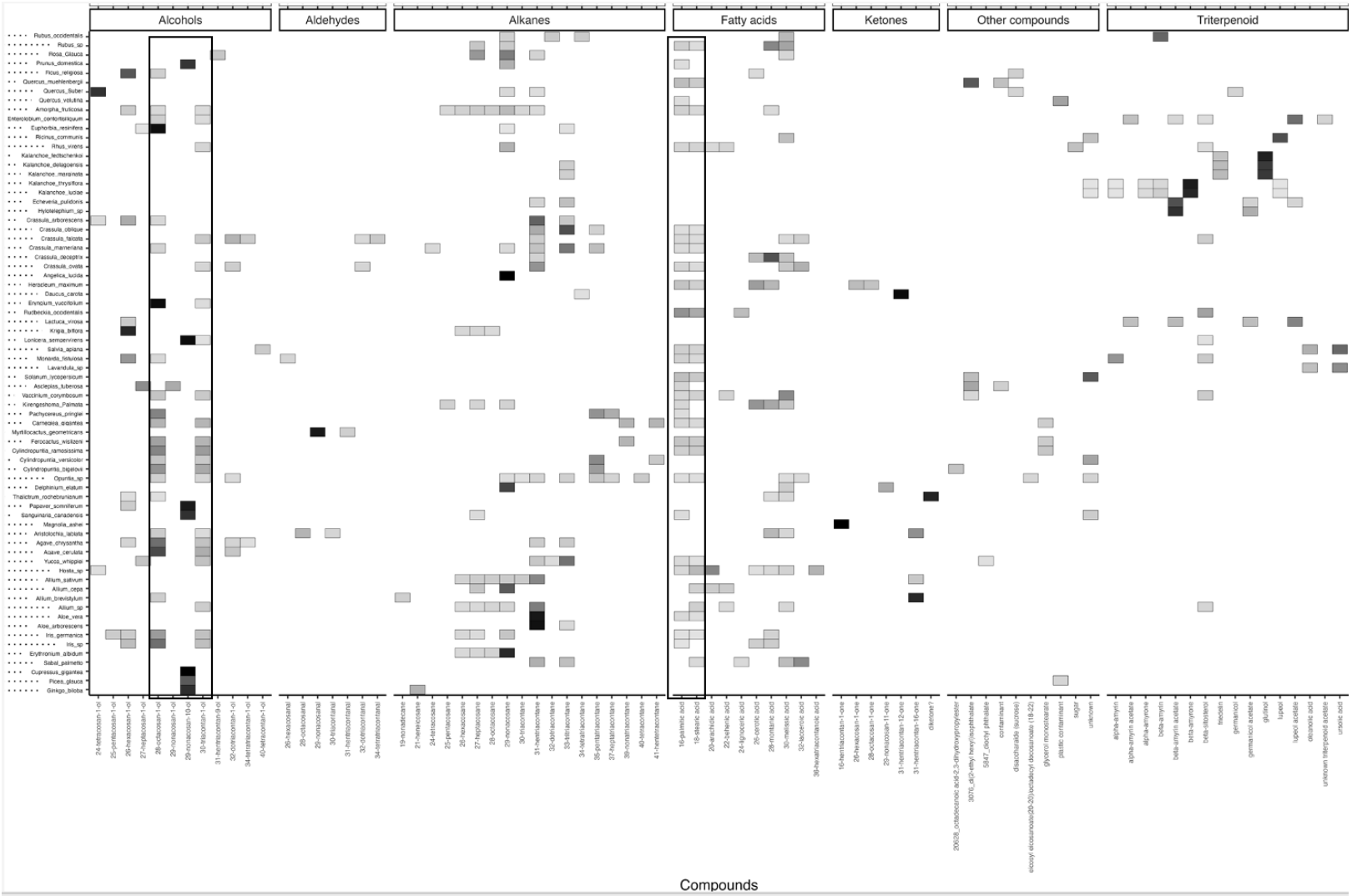

Figure 14. The presence of fatty alcohols 28-octacosanol and 30-triacontanol and fatty acids 16-palmitic acid and 18-stearic acid are shown in the boxed parts.

3.2.2 Within each group, some contain a few extremely common compounds; but other groups contain multiple compounds with more equal number of appearances in various species

Observing each compound group in Figure 15, fatty acid is the most common group found in all the analyzed plant species. This group is represented by ten different fatty acids, ranging from 16-palmitic acid to 36-hexacontanoic acid. However, it is also noticeable that there is an imbalance in how often each of these fatty acids are found in the analyzed species: 16-palmitic acid and 18-stearic acid absolutely have the most predominant occurrences with appearances in 29 and 26 plant species, respectively; and 30-melissic acid comes at a distant third with appearance in 16 different species; while all other fatty acid appear in ten or fewer species. Similar to the fatty acid group, the fatty alcohol group is represented by 12 different compounds and seems to have an uneven trend as well in how often each of them appears in the analyzed plant species. 28-octacosanol and 30-triacontanol are the most common with occurrences in 23 and 21 plant species, followed by 26-hexacosanol appearing in 11 plant species. Meanwhile, each of the other nine alcohols appear in fewer than seven species.

In contrast to fatty acids and alcohols, triterpenoids and alkanes show a more spread-out distribution in how often each of their compounds appears in a number of plant species. In the alkane group, four out of the total 18 compounds (27-heptacosane, 29-nonacosane, 31-hentriacontane and 33-tritriacontane) occur in more than ten species and another three (26-hexacosane, 28-octacosane and 35-pentatriacontane) appear in at least five species.

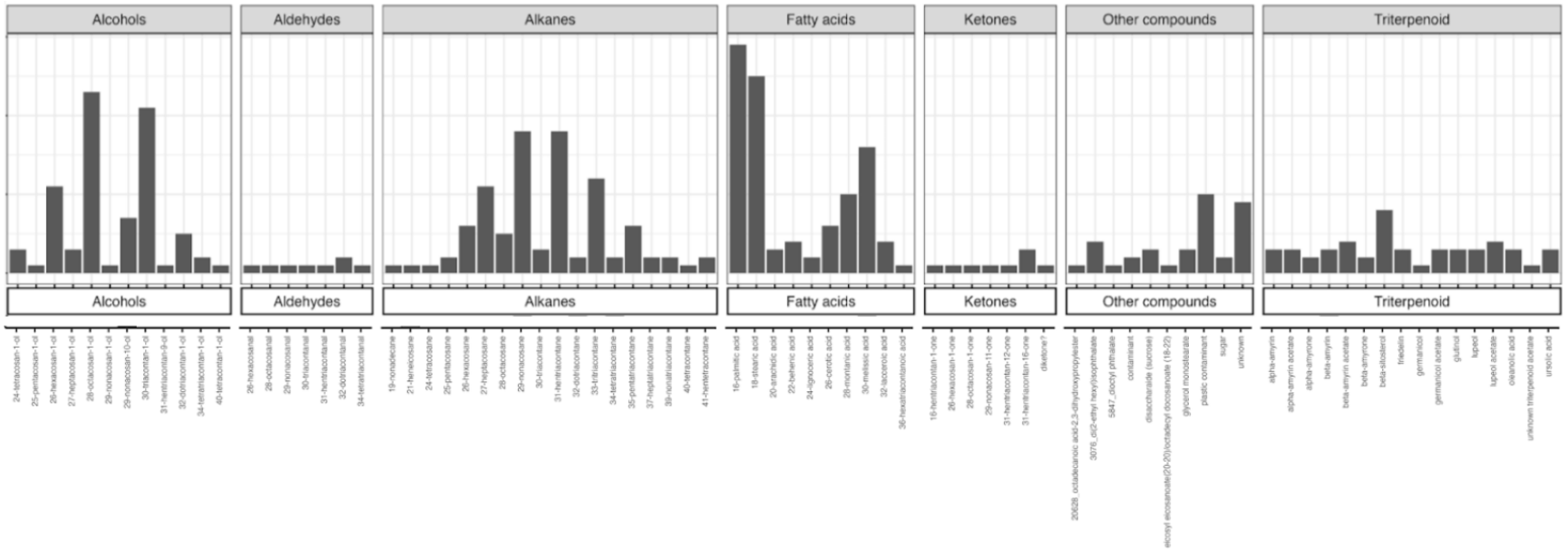

Figure 15. A compact figure of each compound by the number of their occurrences in each of the seven groups. Each tick on the y-axis indicates occurrences in five different species.

The even distribution of each triterpenoid is even more pronounced within the group. Of the 16 reported triterpenoids, only one compound (beta-sitosterol) appears in six species, while no triterpenoids other than beta-sitosterol appear in more than four plant species. This is unlike fatty acid and alcohol in which a few compounds in the group make up the majority of occurrences across a number of plant species.

A possible explanation for this trend of equal occurrence number is that the functions of alkane and triterpenoid and structure of triterpenoids are more complex, and each compound can all be used for specific purposes in plant function. That's why plants produce all those compounds equally and do not just focus on a few favorable structures. Another possible reason is that some of the samples were being contaminated, as 16-palmitic acid and 18-stearic acids are present on human skins' surface; although this seems unlikely to happen to almost half the plant species being analyzed. In addition, it is interesting to point out that this even distribution is very contrasting to what was observed for triterpenoids in Chapter 2. In Chapter 2, a few triterpenoids (alpha-amyrin, beta-amyrin, oleanolic acid, ursolic acid and lupeol) are extremely common in many species in comparison to other documented triterpenoids; while in this map from collected samples indicate that each triterpenoid are very equally presented by the number of plant species containing them.

3.2.3 Considering fatty alcohols with the same chain length, secondary alcohols seem to be more commonly detected than primary alcohols

In the fatty alcohol group, while 11 out of the 12 recorded alcohols are primary alcohols, the only secondary alcohol 29-nonacosan-10-ol seems to be the more common form than the equivalent primary alcohol 29-nonaconsan-1-ol (see Figure 13). A possible explanation for this trend is through the biosynthetic scheme in plant wax, in which primary and secondary fatty alcohols are derived. In this pathway, even-chain primary alcohols are synthesized from even-chain fatty acids by the enzyme CER4 (Rowland et al. 2006; Yeats at al. 2013); while the same even-chain fatty acids are being converted into odd-chain alkanes by the enzyme CER1 and CER3 (Yeats at al. 2013) and then these alkanes are converted into odd-chain secondary alcohols by the enzyme MAH1 (Greer et al. 2007; Yeats at al. 2013). As 29-nonacosanol is an odd-chain fatty alcohol, it is more likely to be synthesized in the form of secondary alcohols according to this reaction scheme.

3.2.4 Alkane chain lengths are normally distributed

Interestingly, the distribution of each alkane within the group by their occurrences seems to show a unique Gaussian trend (see Figure 15). Of the alkanes ranging from 19-nonadecane to 41-hentetracontane, the two most common compounds 29-nonacosane and 31-hentriacontane appears in 18 plant species; following by 27-heptacosane and 33-tritriacontane which appear in 11 and 12 species, respectively; and by 26-hexadecane and

35-pentatriacontane which appear in six species; and other alkanes outside that range only appear in one or two species. This fascinating trend can be interpreted through the reaction scheme of plant chemistry: it has been stated in the previous section that fatty acid chains are the precursor for both alkanes and fatty alcohols with an n-carbon chain of fatty acid producing an n-1-carbon chain of alkane (Zhou et al. 2016). The enzyme CER6 and CER2 plays a role in elongation synthesis of 30-carbon fatty acids (Haslam et al. 2012; Pascal et al. 2013; Yeats at al. 2013) while CER26 helps in synthesis of 32-carbon fatty acids (Pascal et al. 2013; Yeats at al. 2013); and these 30-carbon fatty acids and 32-carbon fatty acids are then respectively converted into 29-nonacosane and 31-hentriacontane by the combined presence of enzymes CER3 and CER1 (Bernard et al. 2012; Yeats at al. 2013). It is possible that these enzymes CER6, CER2 and CER26 combinations produce a normal distribution during the elongation and synthesis of the supposed fatty acid and fatty alkane chains, so some shorter and longer chains than 29-nonacosane and 31-hentriacontane were being created. This explains the Gaussian distribution of each alkane length by the appearances in a number of plant species with 29-nonacosane and 31-hentriacontane being the most common.

It is also necessary to acknowledge that this particular trend of alkanes in the map also helps to support some established information from literature: long chain alkanes (C21-C37) have long been observed as important parts and been utilized as biomarkers of terrestrial plant wax (Bush et al. 2013), which all samples of this project originate from. Specifically, like the map indicates, 29-nonacosane and 31-hentriacontane are very common alkanes being observed in these terrestrial plants (Kawamura et al. 2003; Leider

et al. 2013), where 29-nonacosane is documented to appear in woody plants and 31-hentriacontane is documented to appear in graminoids (grasses) (Bush et al. 2013).

3.2.5 Occurrences of triterpenoid alcohols vs. triterpenoid ketones vs. triterpenoid acetates

From the 376 compound occurrences in the map in Figure 13, triterpenoids occur in 51 of these occurrences, so it can be said that triterpenoids do not make up the majority of occurrences among the surface wax compounds within a realistic and non-biased set of naturally collected plant samples. Moreover, it is found among the 51 triterpenoids occurrences: 32 are triterpenoid alcohols, 4 are triterpenoid ketones, 15 are triterpenoid acetate. The statistical ratio between triterpenoid alcohol: triterpenoid ketone: triterpenoid acetate is 2.1:0.27:1. When being compared with the ratio of 16.6:2.7:1 from the documented set of triterpenoids in Chapter 2, it is clear the triterpenoid acetates are a lot more common in the experimental samples than triterpenoids ketones while it is the opposite from the data set of published studies. On the other hand, triterpenoid alcohols are the most common triterpenoids in both data sets, and they make up 60% and 62.7% of all triterpenoid occurrences in the literature data and the experiment data, respectively.

3.2.6 The presence of fatty alcohols and triterpenoids on the maps show a mutually exclusive trend

Observing Figure 16, while fatty alcohol can be found across the heat map, there seem to be more clusters of concentration of alcohol compounds at the bottom half of the map. In contrast to alcohols, triterpenoids are observed to be concentrated in the top half of the map, with almost no triterpenoids appearing in plant species at the bottom. A possible explanation for this trend is that alcohol and triterpenoids might perform similar roles in plant functions, and thus plants produce them in a way that they are mutually exclusive to each other. However, this explanation has not been documented from previous literature so this can be a potential hypothesis to be tested in future experiments.

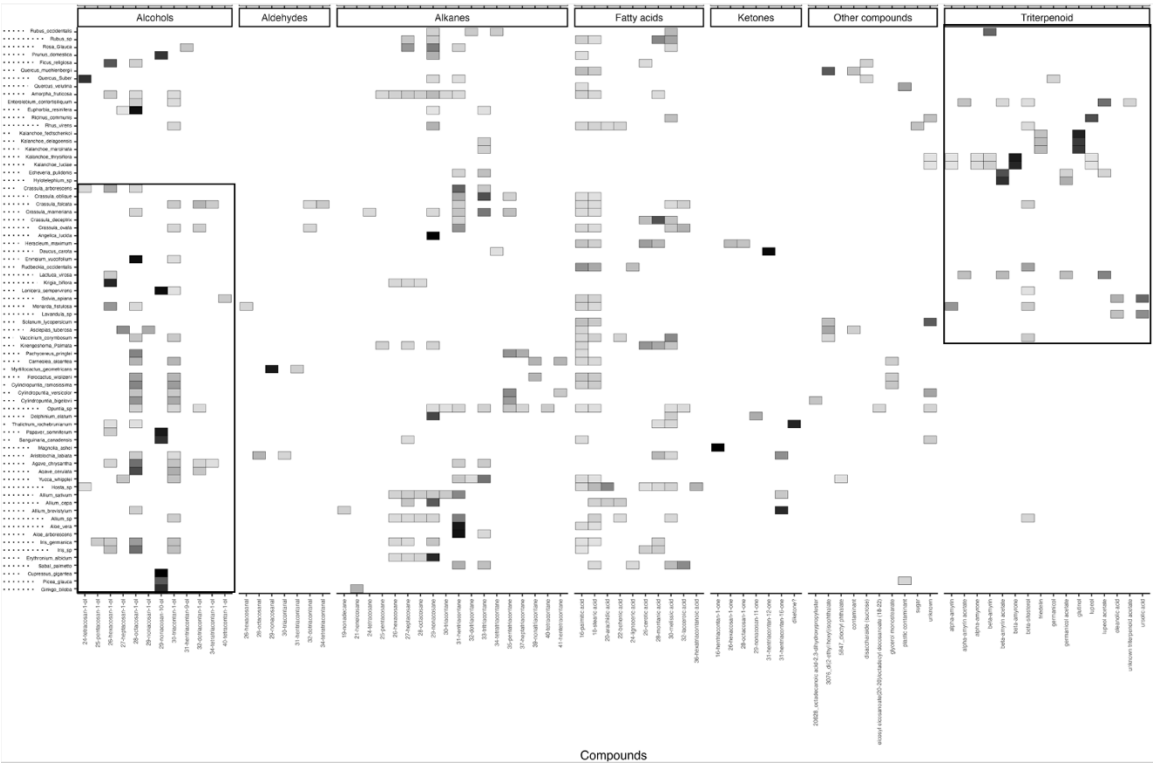

Figure 16. Clusters of fatty alcohol and triterpenoid occurrences are shown in two boxed areas to demonstrate their mutual exclusivity.

3.2.7 Ketones & aldehydes & other compounds

Each of the ketone, aldehyde and other compounds do not have occurrences in more than three plant species so the information is insufficient for any trends to be observed.

3.2.8 Divergent evolution trend

a) Closely related species are more likely to contain similar chemical compounds

In the maps (see Figure 13), a group of closely related species including *Carnegiea gigantea*, *Ferocactus wislizeni*, *Cylindropuntia* species (*C. ramosissima*, *C. versicolor* and *C. bigelovii)* and *Opuntia* species all contain both 28-octacosanol and 30-triacontanol with significant amount (see Figure 17). Another group of related species *Agave chrysantha*, *Agave cerulata* and *Yucca whiplei* also have 28-octacosanol and 30-triacontanol (*Y. whiplei* only has 30-triacontanol) and this group is not very far from the previous group on the phylogenetic tree. For species containing triterpenoids, a group of closely related *Kalanchoe* species (*K. fedtschenkoi*, *K. delagoensis* and *K. marginata*) share glutinol as a major compound and friedlin as a more minor compound. Besides that, two *Kalanchoe* species *K. thyrsiflora* and *K. luciae* share up to five triterpenoids: beta-amyrone as a major compound and alpha-amyrin, alpha-amyrone, beta-amyrin and lupeol as minor compounds (see Figure 17). The two species also have sugar as a minor part of their composition. Regarding species containing alkanes, a group of closely related *Crassula* species (*C. arborescens*, *C. oblique*, *C. faicata*, *C. marneriana*, *C. deceptrix* and *C. ovata*) all share 31-hentriacontane as their major component. Four of the species (with

the exception of *C. arborescens* and *C. deceptrix*) also have both 16-palmitic acid and 18-stearic acid as minor components.

Consequently, all these trends point to a similar conclusion from the triterpenoid's maps in Chapter 2, which indicates that closely related species are a lot more likely to share many similar chemical compounds. As this map in Chapter 3 also contains other compounds besides triterpenoids, it helps to confirm that the previously mentioned trait is not just limited to triterpenoids. And similar to the explanation in Chapter 2, it seems clear that these phylogenetically close species share a common ancestor in the past, explaining why they contain many common compounds. This fits perfectly to the trend of showing divergent evolution traits between the plant species.

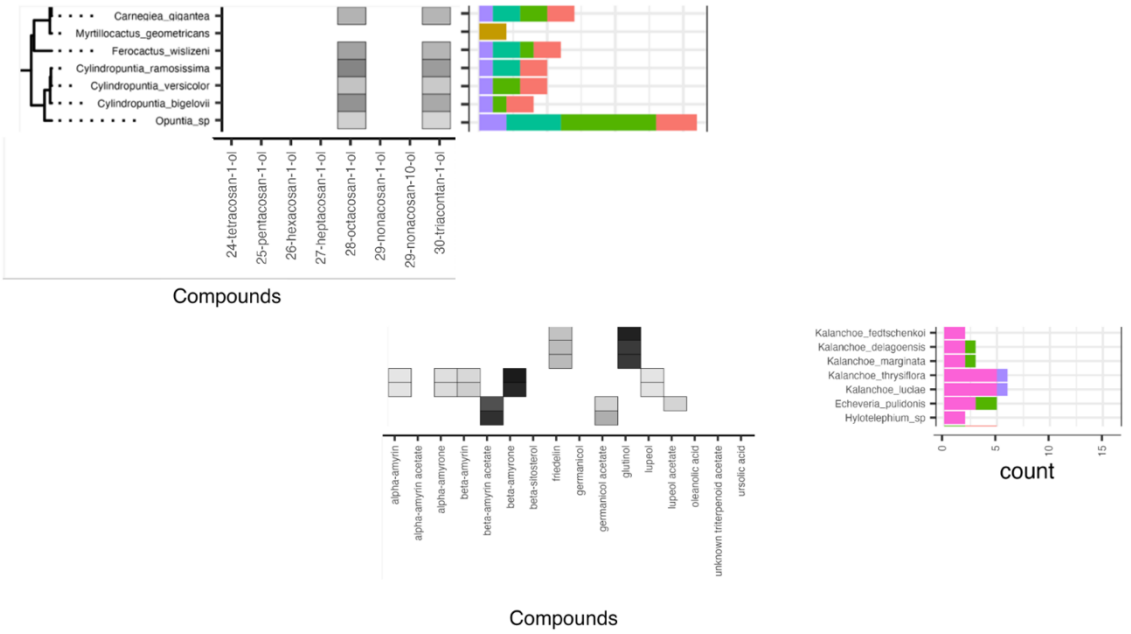

Figure 17. A visible trait reflecting divergent evolution is observed among triterpenoids in the *Kalanchoe* and the *Cylindropuntia* species.

b) Closely related species do not necessarily contain similar number of compounds (diversity level)

Observing Figure 17 and the bar plot from Figure 13, all the *Kalanchoe* species (*K. fedtschenkoi*, *K. delagoensis*, *K. marginata*, *K. thyrsiflora* and *K. luciae*) and the closely related species *Echeveria pulidonis* contains a fairly similar number of two to five total compounds and numbers of compounds within each group: two to five triterpenoids and zero to two alkanes. Similarly, the four closely related species *Pachycereus pringlei*, *Carnegiea gigantea*, *Ferocactus wislizeni*, *Cylindropuntia ramosissima* all have a very similar range of compound diversity of one to two fatty acids, zero to two alkanes and one to two alcohols. In addition to that, the three closely related *Allium sativum*, *Allium cepa* and *Allium brevistylum* have very different number of total compound and the number of compounds in each group: *A. sativum* has six alkanes, *A. cepa* has three fatty acids and two alkanes and *A. brevistylum* has one alkane and one alcohol.

As a result, these cases show that closely related species can have very similar numbers of compound overall and number compound in each group, but that is not always the case. It strengthens a similar conclusion about triterpenoids in Chapter 2, which states that closely related species are more likely to have similar triterpenoids but not necessarily the total number of triterpenoids. This time the conclusion is found to be applied to other compounds beside triterpenoid as well. Furthermore, it helps to enforce the divergent evolution explanation: closely related plant species come from a common ancestor, so they share certain common compounds, but they also have differences in

other compounds when they branched along the evolution path to adapt to different living conditions.

3.2.9 Convergent evolution trend

The results in Chapter 3 also present some features of convergent evolution similar to the species map in Chapter 2. The two very distantly related species *Amorpha fruticosa* and *Allium sativum* share up to six alkane compounds: 26-hexacosane, 27-heptacosane, 28-octacosane, 29-nonacosane, 30-triacontane and 31-hentriacontane (see Figure 18). *Amorpha fruticosa* also shares up to four alkanes compounds with the very distantly related species *Erythronium albidum* : 26-hexacosane, 27-heptacosane, 28-octacosane and 29-nonacosane as major compounds.

Therefore, although it is common to see similar compounds being shared by closely related species, these cases prove that very distantly related species can sometimes contain multiple similar compounds which their closely related species do not even have. This certainly reflects a trend of convergent evolution: plant species without common ancestor can start to produce certain compounds to adjust their biochemistry in a similar fashion to better survive in the same environment. This supports the same conclusion about convergent evolution of triterpenoids observed in Chapter 2, only this time it is observed in alkanes instead.

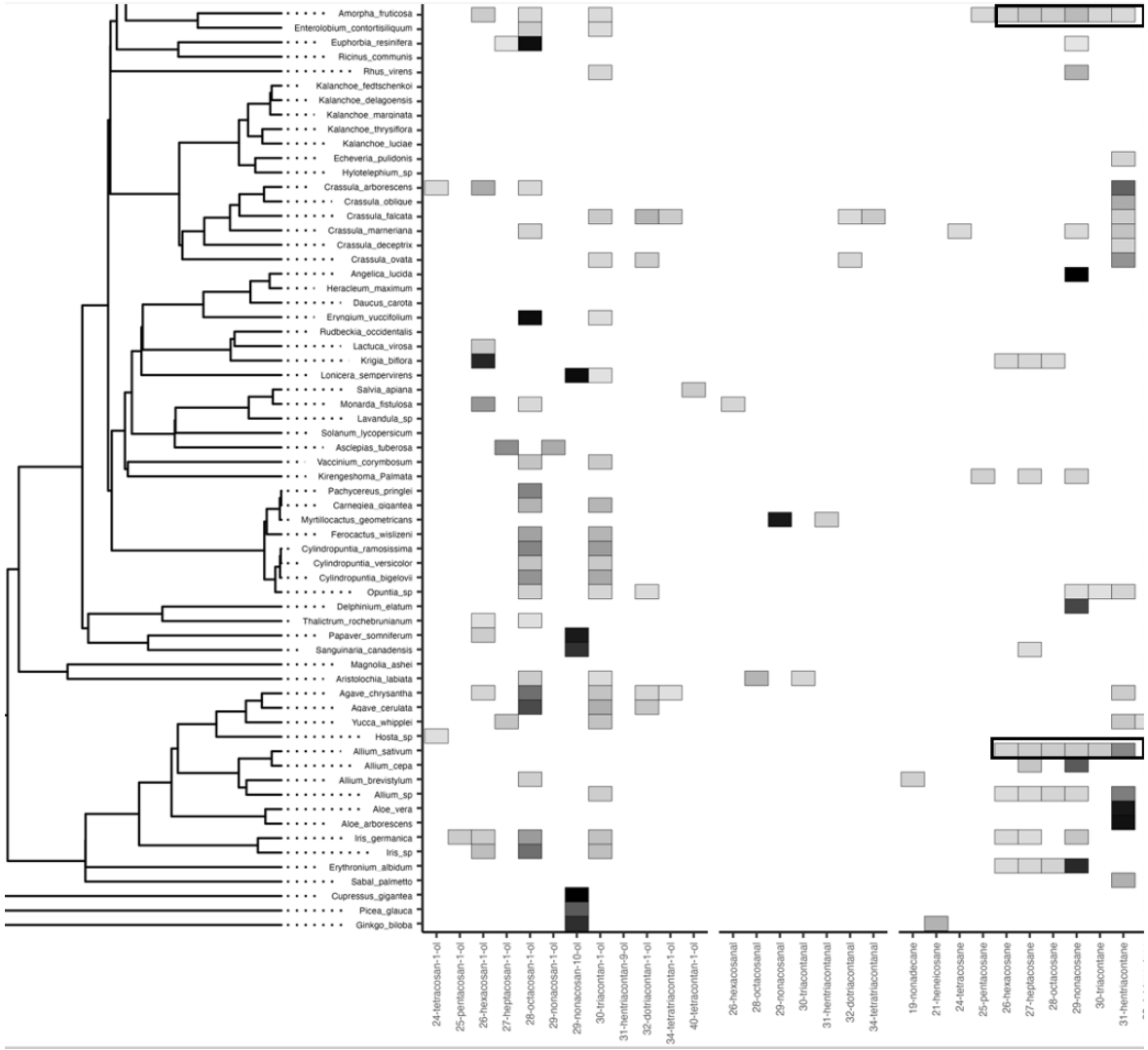

Figure 18. A visible trait reflecting convergent evolution is observed among triterpenoids between the two species *Amorpha fruticosa* and *Allium sativum*

5. Conclusion

The map of compound occurrences in various plant species' surface wax described in this chapter was built based on 130 experimental samples and enabled many analytical interpretations of triterpenoids with other compounds present as well as their interactions with each other. Like in the literature-based map in Chapter 2, the wax bloom-based map in Chapter 3 also contains several compounds like 16-palmitic acid, 18-stearic acid, 28-octacosanol, and 30-triacontanol that appear in close to half of the documented species and are much more common than other compounds. The abundances of these compounds are hypothesized to be due to their importance in plant biochemical pathway and function in general and specific evidence for 16-palmitic acid and 18-stearic acid in this role have also been shown in literature. Moreover, evidences of divergent and convergent traits are also observed in the Chapter 3 map and this really helps to solidify the arguments from the previous chapter. This Chapter 3 map shows that phylogenetically close species are more likely to contain similar chemical compounds including compounds besides triterpenoids, but these closely related species do not necessarily contain similar numbers of compounds from different groups (fatty acid, fatty alcohols, alkanes …) within their wax tissues. This trend demonstrates the divergent evolution traits observed with triterpenoids in Chapter 2, but now it is also confirmed to occur with other compounds: closely related species originate from a common ancestor, so they retain many common compounds, but they also contain different compounds to better adapted to their environment as they branched along different evolution paths. In addition to that, convergent traits are also observed and recorded in Chapter 3: some distantly related

species are shown to have multiple same compounds which are not even shared by their phylogenetic species. This certainly indicates that these distantly related species, without recent common ancestor, possibly start to produce similar compounds so they can function effectively within similar environments. The convergent traits are thus confirmed with several alkene compounds in Chapter 3 in a similar fashion with triterpenoids in Chapter 2. In summary, the data from the wax blooms presented here in Chapter 3 support the conclusion drawn from the literature-based map presented in Chapter 2.

Though these trends of compound presence in plant species all exhibit very solid evidence of divergent and convergent evolutions; there is an issue in that such conclusions rely on the absence of a compound from a particular species. In the datasets explored here, the absence of a compound from a particular species does not guarantee that it was not there or that the species is incapable of producing the compound, only that the investigator did not detect it at the time of analysis. Additionally, some compounds that appear only occasionally may be occasionally detected because they are less stable than the frequently detected compounds. This can be an issue since there is limited control of the quality of the collected samples in Chapter 3, with the way of citizen-scientist sample acquisition. Consequently, additional evidence is necessary to draw firm conclusions about certain evolutionary patterns observed here. This evidence can come from the genetic level: if certain compounds fail to appear in certain species, then in these species the absence of the enzymes that would be required in the final steps of generation

of these absent compounds need to be checked. This is not within the scope of this project, but it is something that is worth looking in for related experiments in the future.

In addition to supporting arguments from Chapter 2, the map in Chapter 3 also reveals some fascinating new trends. As this map contains different compound groups beside triterpenoids, the first noticeable trend being seen is that the differences in distribution of each compound within their respective group: groups like fatty acid and fatty alcohol contain a few very common individual compounds that appear in a lot of plant species, while groups like triterpenoid and alkane include several specific compounds that appear in more equal numbers of plant species. It is thus speculated that the triterpenoid structures, and triterpenoid and alkane functions are generally more complex and have more specific purpose in plant biochemistry. This may explain why different triterpenoids and alkanes are being produced equally in contrast to clear favorability toward a few types of fatty acids and fatty alcohols.

In addition to this, another two special trends were observed that reflected the biosynthesis scheme in plant wax. The first special trend is that considering 29-nonacosanol, its secondary alcohol forms seem to be more common and appear in higher relative abundance than its primary alcohol forms. This is explained by the enzymatic scheme in which even-chain alcohols are produced in primary forms from even-chain fatty acids by enzyme CER4, and odd-chain alcohols are produced in secondary forms from odd-chain alkanes by MAH1 which in turn was made from even-chain fatty acids by enzyme CER1 and CER3. As 29-nonacosanol is an odd-chain fatty alcohol, it makes

sense that it appears more commonly in the secondary forms. A second special trend reflecting the biosynthesis scheme is how each alkane within the compound group shows a distinctive Gaussian trend by their number of occurrences in different plant species: 29-nonacosane and 31-hentriacontane appears in the most species and the longer or shorter the chain of an alkane is from those two alkanes; the fewer plant species that particular alkane appears in. The reaction scheme is that 29-nonacosane and 31-hentriacontane (n-1 chain) are converted respectively from 30-carbon and 32-carbon fatty acids by the combination of enzymes CER3 and CER1; and these 30-carbon and 32-carbon fatty acids are in turn created through elongation by the enzymes CER6 and CER2 (for 30-carbon fatty acids) and CER26 (for 32-carbon fatty acids). As these enzymes CER6, CER2 and CER26 make mistakes during elongation, some shorter and longer chains of fatty acid and alkanes are created; thus a Gaussian trend of species occurrences number is observed among the alkanes with 29-nonacosane and 31-hentriacontane being the most common ones.

A final interesting trend being observed from the wax bloom-based map presented here in Chapter 3 is the seemingly mutual exclusivity between fatty alcohols and triterpenoids: fatty alcohols can be found on the whole maps but there is definitely more concentration of fatty alcohols in plant species at the bottom half of the map, meanwhile triterpenoids are seen to appear in species on the top half of the map. In contrast to the previous trends, unfortunately there is no solid explanation backed by established literature for this particular trend. A possible speculation is that fatty alcohols and triterpenoids may play a similar functional role in plant biochemistry; thus, that explains why plants synthesize

them in mutually exclusive manners, either one but not both. However, this alternative hypothesis is very much not yet being investigated and proven in detail by any publications, so this offers a great window of opportunity for future experiments to be performed and research questions to be answered.

Future Direction

In Chapter 2, observed traits indicating divergent and convergent evolution of triterpenoids in plants raised a question about the relationship between triterpenoids produced by plants and their environmental equation. It seems that plants are very capable of adjusting their triterpenoids presence in order to be more successfully adaptable where they grow. This untested hypothesis concluded from the observation in Chapter 2 provides a perfect opportunity for future experiments to be worked on. Furthermore, the observed traits reflecting divergent and convergent evolution of triterpenoids in both Chapter 2 and 3 and other aliphatic compounds in Chapter 3, still contain a crucial problem. The issue is that the absence of a compound from a particular species does not mean that the compound is not there but only that the compound was not detected: compounds that appear only occasionally or are less stable can still exist. This problem found from the work in Chapter 2 and 3 provides another direction for future work: the enzymes required for production of certain absent compounds can be checked to see if a plant indeed contains these absent compounds or not.

In Chapter 3, it was found that there is a seemingly mutual exclusivity between fatty alcohols and triterpenoids in plants and a possible explanation was given: fatty alcohols and triterpenoids may have similar role in plant biochemistry, so plants synthesize either one but not both. This untested hypothesis also offers another chance for future research and questions to be answered.

BIBLIOGRAPHY

Bernard, A., Domergue, F., Pascal, S., Jetter, R., Renne, C., Faure, J., Haslam, R.P., Napier, J.A., Lessire, R., Joubes, J., **2012.** Reconstitution of plant alkane biosynthesis in yeast demonstrates that Arabidopsis ECERIFERUM1 and ECERIFERUM3 are core components of a very-long-chain alkane synthesis complex. *The Plant Cell*, 24(7), pp. 3106-3118.

Bocquet-Appel, J., **2011.** When the World's Population Took Off: The Springboard of the Neolithic Demographic Transition. *Science,* 333(6042), pp. 560-561.

Burow, G.B., Franks, C.D., Acosta-Martinez, V., Xin, Z., **2009.** Molecular mapping and characterization of BLMC, a locus for proffuse wax (bloom) and enhanced cuticular features of Sorghum (Sorghum biocolor (L.) Moench.) *Theoretical and Applied Genetics,* 118, pp. 423-431.

Bush, R.T., McInerney, F.A., **2013.** Leaf wax *n*-alkane distributions in and across modern plants: Implications for paleoecology and chemotaxonomy. *Geochimica et Cosmochimica Acta,* 117, pp. 161-179.

Busta, L., Schmitz, E., Kosma, D., Schnable, J., Cahoon, E., **2021.** A co-opted steroid synthesis gene, maintained in sorghum but not maize, is associated with a divergence in leaf wax chemistry. *Proceedings of the National Academy of Sciences,* 118(12).

De Geyter, E., Smagghe, G., Rahbé, Y. and Geelen, D., **2011.** Triterpene saponins of Quillaja saponaria show strong aphicidal and deterrent activity against the pea aphid Acyrthosiphon pisum. *Pest management science*, *68*(2), pp.164-169.

De Geyter, E., **2012.** *Toxicity and mode of action of steroid and terpenoid secondary plant metabolites against economically important pest insects in agriculture* (Doctoral dissertation, Ghent University).

Greer, S., Wen, M., Bird, D., Wu, X., Samuels, L., Kunst, L., Jetter, R., **2007.** The Cytochrome P450 Enzyme CYP96A15 is the Midchain Alkane Hydroxylase Responsible for Formation of Secondary Alcohols and Ketones in Stem Cuticular Wax of Arabidopsis. *Plant Physiology,* 145(3), pp. 653-667.

Han, N., Bakovic, M., **2015.** Biological Active Triterpenoids and Their Cardioprotective and Anti-Inflammatory Effects. *Journal of Bioanalysis & Biomedicine,* 12(5), pp. 1-11.

Haslam, T.M., Fernandez, A.M., Zhao, L., Kunst, L., **2012.** Arabidopsis ECERIFERUM2 is a component of the fatty acid elongation machinery required for fatty acid extension to exceptional lengths. *Plant Physiology,* 160(3), pp. 1164-1174.

Haslam, T. M., Kunst, L., **2013.** Extending the story of very-long-chain fatty acid elongation. *Plant Science,* 210, pp. 93-107

Hazell, P., **2009.** The Asian Green Revolution. *International Food Policy Research Institute*, IFPRI Discussion Paper 00911, pp. 1-31.

Huang, F.Y., Chung, B.Y., Bentley, M.D., Alford, A.R., **1995.** Colorado potato beetle antifeedants by simple modification of the birch bark triterpene betulin. *Journal of Agricultural and Food Chemistry*, *43*(9), pp.2513-2516.

Kawamura, K., Ishimura, Y., Yamazaki, K., **2003**. Four years' observations of terrestrial lipid class compounds in marine aerosols from western North Pacific. *Global Biogeochemical Cycles,* 17(1), pp. 3-1-3-19.

Larson, G., Piperno, D., Allaby, R., et al., **2014.** Current perspective and the future of domestication studies. *Proceedings of the National Academy of Sciences of the United States of America,* 111(17), pp. 6139-6146.

Leider, A., Hinrichs, K., Schefub, E., Versteegh, G.J., **2013.** Distribution and stable isotopes of plant wax derived *n*-alkanes in lacustrine, fluvial and marine surface sediments along an Eastern Italian transect and their potential to reconstruct the hydrological cycle. *Geochimica et Cosmochimica Acta,* 117, pp. 16-32

Levetin, E., McMahon, K. *Plants and Society;* 8th. Ed., McGraw-Hill Higher Education (International), **2020.**
https://www.mheducation.com/highered/product/plants-society-levetin-mcmahon/M9781259880049.html

Moggia, C., Graell, J., Lara, I., Schmeda-Hirschmann, G., Thomas-Valdés, S. and Lobos, G.A., **2016**. Fruit characteristics and cuticle triterpenes as related to postharvest quality of highbush blueberries. *Scientia Horticulturae*, *211*, pp.449-457.

MacDonald, M., **2014.** K.D.M. Snell. Annals of the Laboring Poor: Social Change and Agrarian England, 1660–1900. (Cambridge Studies in Population, Economy and Society in Past Time.) *Albion,* 19(1), pp. 91–92.

Moiteiro, C., Marcelo Curto, M.J., Mohamed, N., Bailén, M., Martínez-Díaz, R. and González-Coloma, A., **2006.** Biovalorization of friedelane triterpenes derived from cork processing industry byproducts. *Journal of agricultural and food chemistry*, *54*(10), pp.3566-3571.

Nunn, N., Qian, N., **2011.** The Potato's Contribution to Population and Urbanization: Evidence From A Historical Experiment. *The Quarterly Journal of Economics,* 126(2), pp. 593-650.

Pascal, S., Bernard, A., Sorel, M., Pervent, M., Vile, D., Haslam, R.P., Napier, J.A., Lessire, R., Domergue, F., Joubes, J., **2013.** The Arabidopsis cer26 mutant, like the cer2 mutant, is specifically affected in the very long chain fatty acid elongation process. *The Plant Journal,* 73(5), pp. 733-746.

Ritchie, H., Roser, M., **2013.** "Fertilizers". *Published online at OurWorldInData.org.* Retrieved from: https://ourworldindata.org/fertilizers [Online Resource]

Rowland, O., Zheng, H., Hepworth, S.R., Lam, P., Jetter, R., Kunst, L., **2006.** CER4 Encodes an Alcohol-Forming Fatty Acyl-Coenzyme A Reductase Involved in Cuticular Wax Production in Arabidopsis. *Plant Physiology,* 142(3), pp. 866-877.

Roy, A. and Saraf, S., **2006.** Limonoids: overview of significant bioactive triterpenes distributed in plants kingdom. *Biological and Pharmaceutical Bulletin*, *29*(2), pp.191-201.

Schuster, A., Burghardt, M., Alfarhan, A., Bueno, A., Hedrich, R., Leide, J., Thomas, J., Riederer, M., **2016.** Effectiveness of cuticular transpiration barriers in a desert plant at controlling water loss at high temperatures. *AoB Plants,* 8(10).

Singh, B. and Kaur, A., **2018.** Control of insect pests in crop plants and stored food grains using plant saponins: a review. *LWT*, 87, pp.93-101.

Smith, A.F. *The tomato in America: early history, culture and cookery;* University of Carolina Press: Columbia, S.C., **1994;** pp. 195-204.

Smith, K., Evans, D., El-Hiti, G., **2008.** Role of modern chemistry in sustainable arable crop protection. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 363 (1491), pp. 623-637.

Wambugu, F. and Wafula, J. *Advances in maize streak virus disease research in Eastern and Southern Africa;* International Service for the Acquisition of Agri-Biotech Applications: Ithaca, NY, **2000.** https://agris.fao.org/agris-search/search.do?recordID=US201300054282

Yeats, T.H., Rose, J.K.C., **2013.** The Formation and Function of Plant Cuticles. *Plant Physiology,* 163(1), pp. 5-20.

Zhou, Y.J., Buijs, N.A., Zhu, Z., Qin, J., Siewers, V., Nielsen, J., **2016.** Production of fatty acid-derived oleochemicals and biofuels by synthetic yeast cell factories. *Nature Communications,* 7(11709).

Appendix A

```
######## DIEN AIM 1########
installPhylochemistry()

################################################################################
############
################################################################################
############
################################################################################
############
################### Initialization
#############################################################
################################################################################
############
################################################################################
############
################################################################################
############

## Source
        source("http://thebustalab.github.io/phylochemistry/phylochemistry.R")

## Read in the trait data directly from google docs
  occurrences0 <- read_sheet("https://docs.google.com/spreadsheets/d/12UXzZ-
XRXctWS8d6rMOJR5nB0t513_c_j2LEXbK4g5k/edit#gid=0")
  occurrences <- occurrences0[occurrences0$Included == TRUE,]
  occurrences$genus_species <- paste(occurrences$genus, occurrences$species, sep =
"_")

## Compounds without common names get common named their systematic names
  indeces <- which(is.na(occurrences$compound_common_name))
  occurrences$compound_common_name[indeces] <-
occurrences$compound_systematic_name[indeces]

## Assign compound class information to the trait data
  just_compounds <- read_sheet("https://docs.google.com/spreadsheets/d/12UXzZ-
XRXctWS8d6rMOJR5nB0t513_c_j2LEXbK4g5k/edit#gid=0", range =
"just_compounds")

## Check for name mismatches

occurrences$compound_common_name[is.na(match(occurrences$compound_common_n
ame, just_compounds$compound_name))]

## Add compound classes to the occurrence data
```

```
  occurrences$compound_class <-
just_compounds$compound_class[match(occurrences$compound_common_name,
just_compounds$compound_name)]

## filter out occurrences for which we don't know the structure
  occurrences <- occurrences[!occurrences$compound_class == "NA",]
  occurrences <- occurrences[!is.na(occurrences$compound_class),]
  occurrences <- filter(occurrences, compound_relative_abundance != "none detected")


## Build a data frame that describes the species in the dataset
          genus_species <- data.frame(
                    genus_species = unique(paste(occurrences$genus, occurrences$species,
sep = "_")),
                    genus = NA,
                    species = NA
          )
          genus_species$genus <- gsub("_.*$", "", genus_species$genus_species)
          genus_species$species <- gsub(".*_", "", genus_species$genus_species)
  genera <- genus_species[!duplicated(genus_species$genus),]



################################################################################
############
################################################################################
############
################################################################################
############
################### Analysis on the Genus species level
#####################################
################################################################################
############
################################################################################
############
################################################################################
############

## Analysis on the Genus_species level
          ## Build a tree of the genus_species in your dataset
  species_tree <- fortify(buildTree(
    scaffold_type = "newick",
    scaffold_in_path = "http://thebustalab.github.io/data/angiosperms.newick",
    members = unique(paste(occurrences$genus, occurrences$species, sep = "_"))
  ))
```

```
 species_tree <- full_join(species_tree, genus_species, by = c("label" =
"genus_species"))


 ## Draw phylogenetic tree
 phylo_species_tree <- ggtree(species_tree) +
  geom_tiplab(
   aes(label = label), size = 1.8, align = TRUE,
   hjust = 1, offset = 42, geom = "label", label.size = 0
  ) +
  coord_cartesian(xlim = c(0,235))+
  scale_y_continuous(expand = c(0,0.5)) +
  theme_void() +
  theme(plot.margin = unit(c(0.6,0,0,0), "cm"))
 phylo_species_tree


        ## Make the order of species in "occurrences" match the order of the tree
                genus_species_in_this_tree <- filter(species_tree, isTip == TRUE)$label
                genus_species_in_this_tree_ordered <-
genus_species_in_this_tree[order(filter(species_tree, isTip == TRUE)$y, decreasing =
TRUE)]
                occurrences$genus <- factor(occurrences$genus_species, levels =
rev(genus_species_in_this_tree_ordered))
                occurrences <- filter(occurrences, genus_species != "<NA>")
                #occurrences$genus_species

        ## Set color scheme
                colors <- c("red", "pink", "grey")
                names(colors) <- c("major", "minor", "none detected")

        ## Draw a heat map of the compound occurrences
                #occurrences <- occurrences[!is.na(occurrences$genus),]
                #occurrences <-
occurrences[!is.na(occurrences$compound_common_name),]
                yline_positions <- as.numeric(occurrences$genus) + 0.5
                xline_positions <-
as.numeric(factor(occurrences$compound_common_name)) + 0.5
                heat_map_species <- ggplot() +
                        geom_tile(
                          data = occurrences,
                          aes(x = compound_common_name, y = genus, fill =
compound_relative_abundance),
                          color = "black"
                        ) +
```

```
                #geom_hline(aes(yintercept = yline_positions), color = "black", size =
0.05) +
                #geom_vline(aes(xintercept = xline_positions), color = "black", size =
0.05) +
                scale_fill_manual(values = colors, na.value = "grey") +
                scale_y_discrete(name = "") +
                theme_classic() +
                facet_grid(.~compound_class, scales = "free_x", space = "free_x") +
                  theme(
                    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size
= 5),
                    axis.text.y = element_blank(),
                    plot.margin = unit(c(0,0,0,-0.55), "cm")
                    #panel.grid.minor = element_line(color = "black", size = 1)
                  )
              heat_map_species

        ## Draw a bar plots triterpenoid occurences for each species
              ## Make the order of species in "right_plot" match the order of the tree
              occurrences$compound_class <-
just_compounds$compound_class[match(occurrences$compound_common_name,
just_compounds$compound_name)]
              occurrences$genus_species <- factor(occurrences$genus_species, levels =
rev(genus_species_in_this_tree_ordered))
         ## Make the "right_plot"
              occurrences %>%
                dplyr::select(genus_species, compound_common_name) %>%
                unique() %>%
                ggplot() +
                geom_bar(aes(y = genus_species))+
                theme_bw()+
              theme(
                axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 5),
                axis.text.y = element_text(size = 5),
                plot.margin = unit(c(0.6,0,0,-0.5), "cm") )  -> right_plot_species
                #panel.grid.minor = element_line(color = "black", size = 1)
              right_plot_species
              ## Make the "top_plot"
              top_plot_species <- ggplot(occurrences) +
                geom_bar(aes(x = compound_common_name)) +
                theme_bw()+
                theme(
                  axis.text.x = element_blank(),
                  axis.text.y = element_blank(),
                  plot.margin = unit(c(0,0,-0.5,-0.6), "cm")
                  #panel.grid.minor = element_line(color = "black", size = 1)
```

```
  )+
  facet_grid(.~compound_class, space = "free_x", scales = "free_x")
top_plot_species

#ggsave(
  #filename = "/Users/dienngyen/Desktop/test1.png", plot = top_plot ,
  #device = "png", width = 20, height = 12, units = "in", dpi = 600
#)

blank <- ggplot() + theme_void()

#final_plot <- cowplot::plot_grid(
#phylo_species_tree, blank, heat_map_species, blank, right_plot,
#align = "h", axis = "b",
#rel_widths = c(0.5,-0.005, 1.5,-0.25, 0.5),
#nrow = 1
#)

col1 <- cowplot::align_plots(blank, phylo_species_tree, align = "v", axis = "l")
col2 <- cowplot::align_plots(blank, blank, align = "v", axis = "l")
col3 <- cowplot::align_plots(top_plot_species, heat_map_species, align = "v",
axis = "r")
col3 <- cowplot::align_plots(col3[[1]], col3[[2]], align = "v", axis = "l")
col4 <- cowplot::align_plots(blank, blank, align = "v", axis = "l")
col5 <- cowplot::align_plots(blank, right_plot_species, align = "v", axis = "l")

bottom_row <- cowplot::align_plots(col1[[2]], col2[[2]], col3[[2]], col4[[2]],
col5[[2]], align = "h", axis = "b")

final_plot <- cowplot::plot_grid(
  col1[[1]], col2[[1]], col3[[1]], col4[[1]], col5[[1]],
  bottom_row[[1]], bottom_row[[2]], bottom_row[[3]], bottom_row[[4]],
bottom_row[[5]],
  # align = "h", axis = "b",
  rel_widths = c(0.6, -0.03, 1.5,-0.225, 0.4, 0.6, -0.03, 1.5,-0.225, 0.4),
  nrow = 2,
  rel_heights = c(1,4)
)

ggsave(
  filename = "/Users/dienngyen/Desktop/test.png", plot = final_plot,
   device = "png", width = 22, height = 15, units = "in", dpi = 600
)

ggsave(
  filename = "/Users/dienngyen/Desktop/test1.png", plot = final_plot,
```

```
  device = "png", width = 22, height = 15, units = "in", dpi = 600
)


##  Analysis with just "yes" and "no" on the Genus_species level
## Build a tree of the genus_species in your dataset
species_tree <- fortify(buildTree(
  scaffold_type = "newick",
  scaffold_in_path = "http://thebustalab.github.io/data/angiosperms.newick",
  members = unique(paste(occurrences$genus, occurrences$species, sep = "_"))
))
species_tree <- full_join(species_tree, genus_species, by = c("label" =
"genus_species"))

## Draw phylogenetic tree
phylo_species_tree <- ggtree(species_tree) +
  geom_tiplab(
    aes(label = label), size = 1.8, align = TRUE,
    hjust = 1, offset = 50, geom = "label", label.size = 0
  ) +
  coord_cartesian(xlim = c(0,235))+
  scale_y_continuous(expand = c(0,0.5)) +
  theme_void() +
  theme(plot.margin = unit(c(0,0,0,0), "cm"))
phylo_species_tree

## Make the order of species in "occurrences" match the order of the tree
genus_species_in_this_tree <- filter(species_tree, isTip == TRUE)$label
genus_species_in_this_tree_ordered <-
genus_species_in_this_tree[order(filter(species_tree, isTip == TRUE)$y, decreasing =
TRUE)]
occurrences$genus <- factor(occurrences$genus_species, levels =
rev(genus_species_in_this_tree_ordered))

## Set color scheme
colors <- c("red", "red", "grey")
names(colors) <- c("major", "minor", "none detected")

## Draw a heat map of the compound occurrences
occurrences <- occurrences[!is.na(occurrences$genus),]
occurrences <- occurrences[!is.na(occurrences$compound_common_name),]
yline_positions <- as.numeric(occurrences$genus) + 0.5
xline_positions <- as.numeric(factor(occurrences$compound_common_name)) +
0.5
heat_map_species <- ggplot() +
  geom_tile(
```

```
            data = occurrences,
            aes(x = compound_common_name, y = genus, fill =
compound_relative_abundance),
            color = "black"
          ) +
          geom_hline(aes(yintercept = yline_positions), color = "black", size = 0.05) +
          geom_vline(aes(xintercept = xline_positions), color = "black", size = 0.05) +
          scale_fill_manual(values = colors, na.value = "grey") +
          scale_y_discrete(name = "") +
          theme_classic() +
          theme(
            axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 5),
            axis.text.y = element_blank(),
            plot.margin = unit(c(0,0,0,-0.6), "cm")
            #panel.grid.minor = element_line(color = "black", size = 1)
          )
        heat_map_species

        blank <- ggplot()

        final_plot <- cowplot::plot_grid(
          phylo_species_tree, blank, heat_map_species,
          align = "h", axis = "b",
          rel_widths = c(0.5, -0.005, 1.5),
          nrow = 1
        )
        final_plot

        ggsave(
          filename = "/Users/dienngyen/Desktop/test.png", plot = final_plot,
          device = "png", width = 20, height = 12, units = "in", dpi = 600
        )


################################################################################
############
################################################################################
############
################################################################################
############
################### Checking for contradicting information from different sources ###
################################################################################
############
################################################################################
############
```

```
##############################################################################
############

occurrences <- read_sheet("https://docs.google.com/spreadsheets/d/12UXzZ-
XRXctWS8d6rMOJR5nB0t513_c_j2LEXbK4g5k/edit#gid=0")
all_reports <- paste(    occurrences$genus,    occurrences$species,
          occurrences$reference)
all_reports <- unique(all_reports)
all_reports[order(all_reports)]

reports <- paste(         occurrences$genus,    occurrences$species)
reports <- unique(reports)
reports[order(reports)]

reports <- paste(occurrences$extract, occurrences$genus,    occurrences$species)
reports <- unique(reports)
reports[order(reports)]

table(occurrences$extract)




##############################################################################
############
##############################################################################
############
##############################################################################
############
################## Analysis on the Genus level
#############################################
##############################################################################
############
##############################################################################
############
##############################################################################
############

          ## Build a tree of the genus_species in your dataset
            genus_tree <- fortify(buildTree(
                              scaffold_type = "newick",
                              scaffold_in_path =
"http://thebustalab.github.io/data/angiosperms.newick",
                              members = genera$genus_species,
                    ))
```

```
        genus_tree <- full_join(genus_tree, genera, by = c("label" =
"genus_species"))

    ## Draw phylogenetic tree
     #      phylo_genus_tree <- ggtree(genus_tree) +
            #       geom_tiplab(aes(label = genus)) +
            #       theme_classic() +
            #       coord_cartesian(xlim = c(0,300))
            #phylo_genus_tree

    ## Draw phylogenetic tree
            phylo_genus_tree <- ggtree(genus_tree) +
              geom_tiplab(
                aes(label = genus), size = 1.8, align = TRUE,
                hjust = 1, offset = 50, geom = "label", label.size = 0
              ) +
              coord_cartesian(xlim = c(0,420))+
              scale_y_continuous(expand = c(0,0.5)) +
              theme_void() +
              theme(plot.margin = unit(c(0.65,0,1,0), "cm"))
            phylo_genus_tree

    ## Make the order of genus in "occurrences" match the order of the tree
            genus_species_in_this_tree <- filter(genus_tree, isTip == TRUE)$label
            genus_species_in_this_tree_ordered <-
genus_species_in_this_tree[order(filter(genus_tree, isTip == TRUE)$y, decreasing =
TRUE)]

            genus_in_this_tree <- filter(genus_tree, isTip == TRUE)$genus
            genus_in_this_tree_ordered <- genus_in_this_tree[order(filter(genus_tree,
isTip == TRUE)$y, decreasing = TRUE)]

            occurrences$genus <- factor(occurrences$genus, levels =
rev(genus_in_this_tree_ordered))
            occurrences <- filter(occurrences, genus != "<NA>")
            #occurrences$genus_species

    ## Draw a heat map of the compound occurrences
            #heat_map <- ggplot(data = occurrences) +
            #       geom_tile(aes(x = compound_common_name, y = genus, fill =
compound_relative_abundance)) +
            #  theme(axis.text.x = element_text(angle = 90, hjust = 1, size =5))
            #heat_map

    ## Set color scheme
            colors <- c("red", "pink", "grey")
```

```
                names(colors) <- c("major", "minor", "none detected")

 ## Draw a heat map of the compound occurrences
                #occurrences <- occurrences[!is.na(occurrences$genus),]
                #occurrences <-
occurrences[!is.na(occurrences$compound_common_name),]
                yline_positions <- as.numeric(occurrences$genus) + 0.5
                xline_positions <-
as.numeric(factor(occurrences$compound_common_name)) + 0.5
                heat_map_genus <- ggplot(data = occurrences) +
                  geom_tile(
                    data = occurrences,
                    aes(x = compound_common_name, y = genus, fill =
compound_relative_abundance),
                    color = "black"
                  ) +
                  #geom_hline(aes(yintercept = yline_positions), color = "black", size =
0.05) +
                  #geom_vline(aes(xintercept = xline_positions), color = "black", size =
0.05) +
                  scale_fill_manual(values = colors, na.value = "grey") +
                  scale_y_discrete(name = "") +
                  theme_classic() +
                  facet_grid(.~compound_class, scales = "free_x", space = "free_x") +
                  theme(
                    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5,size = 5),
                    axis.text.y = element_blank(),
                    plot.margin = unit(c(0,0,0,-0.65), "cm")
                    # panel.grid.minor = element_line(color = "black", size = 1)
                  )
                heat_map_genus

 ## Draw a bar plots triterpenoid occurrences for each species
                ## Make the order of species in "right_plot" match the order of the tree
                occurrences$compound_class <-
just_compounds$compound_class[match(occurrences$compound_common_name,
just_compounds$compound_name)]
                occurrences$genus <- factor(occurrences$genus, levels =
rev(genus_in_this_tree_ordered))
                ## Make the "right_plot"
                occurrences %>%
                  dplyr::select(genus, compound_common_name) %>%
                  unique() %>%
                  ggplot() +
                  geom_bar(aes(y = genus))+
                  theme_bw()+
```

```
                  theme(
                    axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 5),
                    axis.text.y = element_text(size = 5),
                    plot.margin = unit(c(0.6,0,0,-0.5), "cm") ) -> right_plot_genus
                    #panel.grid.minor = element_line(color = "black", size = 1)
                  right_plot_genus

                  ## Make the "top_plot"
                  occurrences %>%
                    dplyr::select(genus, compound_common_name, compound_class) %>%
                    ggplot() +
                    geom_bar(aes(x = compound_common_name)) +
                    scale_y_continuous(name = "")+
                    theme_bw()+
                    theme(
                      axis.text.x = element_blank(),
                      axis.text.y = element_blank(),
                      plot.margin = unit(c(0,0,-0.5,-0.4), "cm")
                      #panel.grid.minor = element_line(color = "black", size = 1)
                    )+
                    facet_grid(.~compound_class, space = "free_x", scales = "free_x") ->
top_plot_genus
                  top_plot_genus


                  blank <- ggplot() + theme_void()


          col1 <- cowplot::align_plots(blank, phylo_genus_tree, align = "v", axis = "l")
          col2 <- cowplot::align_plots(blank, blank, align = "v", axis = "l")
          col3 <- cowplot::align_plots(top_plot_genus, heat_map_genus, align = "v", axis
= "r")
          col3 <- cowplot::align_plots(col3[[1]], col3[[2]], align = "v", axis = "l")
          col4 <- cowplot::align_plots(blank, blank, align = "v", axis = "l")
          col5 <- cowplot::align_plots(blank, right_plot_genus, align = "v", axis = "l")

          bottom_row <- cowplot::align_plots(col1[[2]], col2[[2]], col3[[2]], col4[[2]],
col5[[2]], align = "h", axis = "b")

          final_plot1 <- cowplot::plot_grid(
            col1[[1]], col2[[1]], col3[[1]], col4[[1]], col5[[1]],
            bottom_row[[1]], bottom_row[[2]], bottom_row[[3]], bottom_row[[4]],
bottom_row[[5]],
            # align = "h", axis = "b",
            rel_widths = c(0.5, -0.221, 1.5,-0.23, 0.5, 0.5, -0.221, 1.5,-0.23, 0.5),
            nrow = 2,
```

```
  rel_heights = c(1,3)
)

        ggsave(
          filename = "/Users/dienngyen/Desktop/test3.png", plot = final_plot1,
          device = "png", width = 20, height = 12, units = "in", dpi = 600
        )
```

Appendix B

```
library(dplyr)
library(ggplot2)
library (tidyverse)
library(ape)
library(googlesheets4)
library(ggtree)

# source("http://thebustalab.github.io/phylochemistry/phylochemistry.R")
source("http://thebustalab.github.io/phylochemistry/integrationAppLite.R")

integrationAppLite("/Volumes/Busta_Lab_2/Susannah_Chase/AIAEXPRT.AIA/350_Ru
n_1")
integrationAppLite("/Volumes/Busta_Lab_2/Susannah_Chase/AIAEXPRT.AIA/350_Ru
n_2")
integrationAppLite("/Volumes/Busta_Lab_2/Susannah_Chase/AIAEXPRT.AIA/350_Ru
n_3")
integrationAppLite("/Volumes/Busta_Lab_2/Susannah_Chase/AIAEXPRT.AIA/450_Ru
n_1")
integrationAppLite("/Volumes/Busta_Lab_2/Susannah_Chase/AIAEXPRT.AIA/450_Ru
n_2")
integrationAppLite("/Volumes/Busta_Lab_2/Susannah_Chase/AIAEXPRT.AIA/450_Ru
n_3")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_1")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_2")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_3")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_5")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_8")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_9")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_11")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_12")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_15")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_16")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_18")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_19")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_20")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_21")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_22")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_23")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_24")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_25")
```

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_26")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_27")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_28")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_29")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_30")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_31")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_32")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_33")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_34")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_35")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_36")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_37")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_38")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_39")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_40")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_41")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_42")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_43")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_44")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_45")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_46")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_47")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_48")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_49")

##########################

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_52")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_53")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_54")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_56")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_62")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_65")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_68")

```
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_74")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_75")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_81")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_82")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_83")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_85",
zoom_and_scroll_rate = 25)
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_88")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_89")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_90")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_91")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_92")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_93")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_94")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_95")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_96")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_97")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_99")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_100")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_103")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_104")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_105")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_106")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_109")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_110")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_111")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_115")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_116")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_117")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_118")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_119")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_120")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_121")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_122")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_124")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_130")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_132")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_138")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_139")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_140")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_141")


integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_142")
```

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_144")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_145")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_147")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_148")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_149")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_150")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_154")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_155")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_159")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_161")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_164")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_169")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_170")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_172")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_180")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_181")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_183")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_184")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_185")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_187")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_194")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_195")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_200")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_204")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_205")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_206")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_207")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_211")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_212")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_217")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_219")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_220")

integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_221")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_222")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_223")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_224")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_225")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_226")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_227")
integrationAppLite("/Volumes/Busta_Lab_2/chemical_blooms/data/PCB_228")

```
#########################################################################
#####################################
#########################################################################
#####################################
############################## MERGING DATA
#########################################################################
#########################################################################
#####################################
#########################################################################
#####################################



#data <- mergePeakLists("/Volumes/Busta_Lab_2/chemical_blooms/data/")
#writeMonolist(data,"/Volumes/Busta_Lab_2/chemical_blooms/data/output.csv")

peak_data <- mergePeakLists("/Volumes/Busta_Lab_2/chemical_blooms/data/")
peak_data$PCB_number <- gsub(".CDF.csv","", gsub(".*/", "",
peak_data$path_to_cdf_csv))

#writeMonolist(peak_data,"/Volumes/Busta_Lab_2/chemical_blooms/data/output.csv")
writeMonolist(peak_data,"/Users/dienngyen/Desktop/output.csv")
PCB_metadata <-
googlesheets4::read_sheet("https://docs.google.com/spreadsheets/d/1gi5VEmVb5CZH7n
ET1uq77k7z2rA_klHDeYx3UYioSgQ/edit#gid=113692019")

chemical_bloom_merged_data <- left_join(peak_data, PCB_metadata, by =
"PCB_number")

chemical_bloom_merged_data %>%
  group_by(PCB_number) %>%
  dplyr::mutate(percent_area = (area/sum(area))*100) -> chemical_bloom_merged_data

select(chemical_bloom_merged_data, Genus, Species, peak_ID, percent_area)
googlesheets4::write_sheet(chemical_bloom_merged_data,
"https://docs.google.com/spreadsheets/d/1Gdrj20dEB2k60BdWC8QD25ls4Ws7hG8IcPo
5CIt4jDk/edit#gid=0", sheet = "chemical_bloom_merged_data")

chemical_bloom_merged_data <-
googlesheets4::read_sheet("https://docs.google.com/spreadsheets/d/1Gdrj20dEB2k60Bd
WC8QD25ls4Ws7hG8IcPo5CIt4jDk/edit#gid=0", sheet =
"chemical_bloom_merged_data")

chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"oleanolic_acid 1"] <- "oleanolic acid"
```

```
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"oleanolic_acid 2"] <- "oleanolic acid"
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"ursolic_acid 1"] <- "ursolic acid"
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"ursolic_acid 2"] <- "ursolic acid"
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"sugar 1"] <- "sugar"
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"sugar 2"] <- "sugar"
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"disaccharaide_1 (sucrose)"] <- "disaccharaide (sucrose)"
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"disaccharaide_2 (sucrose)"] <- "disaccharaide (sucrose)"
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"sucrose"] <- "disaccharaide (sucrose)"
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"disaccharide 1"] <- "disaccharide"
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"disaccharide 2"] <- "disaccharide"
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"disaccharide 3"] <- "disaccharide"
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"triterpenoid"] <- "unknown triterpenoid"
chemical_bloom_merged_data$peak_ID[chemical_bloom_merged_data$peak_ID ==
"triterpenoid acetate"] <- "unknown triterpenoid acetate"

## Create a genus_species
chemical_bloom_merged_data$genus_species <-
paste(chemical_bloom_merged_data$Genus, chemical_bloom_merged_data$Species, sep
= "_")
chemical_bloom_merged_data_unique_PCB <-
chemical_bloom_merged_data[!duplicated(chemical_bloom_merged_data$PCB_number
),]
chemical_bloom_merged_data_unique_PCB$genus_species[duplicated(chemical_bloom
_merged_data_unique_PCB$genus_species)]

## Check on number of unique species at this point (as of 2/16/22 there were 85 species
here)
unique(chemical_bloom_merged_data$genus_species)
unique(chemical_bloom_merged_data$PCB_number)


##########################################################################
#####################################
```

```
##############################################################################
#########################################
############################## FILTERING CONFLICTING INFORMATION
#################################################
##############################################################################
#########################################
##############################################################################
#########################################



chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$Genus == "Unknown",]

## Filter Aloe_Vera
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_144",]
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_181",]
## Filter Allium_sp
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_161",]
## Filter Erythronium_albidum
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_204",]
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_205",]
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_206",]
## Filter Sanguinaria_canadensis
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_211",]
## Filter Kalanchoe_fedtschenkoi
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_227",]
## Filter Crassula_ovata
```

```
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_28",]
## Filter Prunus_domestica
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_33",]
## Filter Rubus_sp
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_74",]
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_90",]
## Filter Tephrocactus_articulatus
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_38",]


## AVERAGE SIMILAR SAMPLES from same species
chemical_bloom_merged_data %>%
  select(Genus, Species, genus_species, PCB_number, peak_ID, percent_area) ->
chemical_bloom_merged_data

## Mean Vaccinium_corymbosum
chemical_bloom_merged_data %>%
  filter(genus_species == "Vaccinium_corymbosum") %>%
  group_by(PCB_number) %>%
  mutate(percent = percent_area/sum(percent_area)) %>% ungroup() %>%
  group_by(peak_ID) %>%
  summarize(percent_area = mean(percent)) %>% ungroup() %>%
  mutate(percent_area = percent_area/sum(percent_area)*100) -> temp

  temp$Genus <- "Vaccinium"
  temp$Species <- "corymbosum"
  temp$genus_species <- "Vaccinium_corymbosum"
  temp$PCB_number <- "NA"
  chemical_bloom_merged_data %>%
          filter(genus_species != "Vaccinium_corymbosum") ->
chemical_bloom_merged_data
  chemical_bloom_merged_data <- rbind(chemical_bloom_merged_data, temp)


## Mean Iris_germanica
  chemical_bloom_merged_data %>%
```

```
        filter(genus_species == "Iris_germanica") %>%
        group_by(PCB_number) %>%
        mutate(percent = percent_area/sum(percent_area)) %>% ungroup() %>%
        group_by(peak_ID) %>%
        summarize(percent_area = mean(percent)) %>% ungroup() %>%
        mutate(percent_area = percent_area/sum(percent_area)*100) -> temp

 temp$Genus <- "Iris"
 temp$Species <- "germanica"
 temp$genus_species <- "Iris_germanica"
 temp$PCB_number <- "NA"
 chemical_bloom_merged_data %>%
        filter(genus_species != "Iris_germanica") -> chemical_bloom_merged_data
 chemical_bloom_merged_data <- rbind(chemical_bloom_merged_data, temp)


## Mean Borodinia_laevigata
chemical_bloom_merged_data %>%
 filter(genus_species == "Borodinia_laevigata") %>%
 group_by(PCB_number) %>%
 mutate(percent = percent_area/sum(percent_area)) %>% ungroup() %>%
 group_by(peak_ID) %>%
 summarize(percent_area = mean(percent)) %>% ungroup() %>%
 mutate(percent_area = percent_area/sum(percent_area)*100) -> temp

temp$Genus <- "Borodinia"
temp$Species <- "laevigata"
temp$genus_species <- "Borodinia_laevigata"
temp$PCB_number <- "NA"
chemical_bloom_merged_data %>%
  filter(genus_species != "Borodinia_laevigata") -> chemical_bloom_merged_data
chemical_bloom_merged_data <- rbind(chemical_bloom_merged_data, temp)


## Mean Kalanchoe_luciae
chemical_bloom_merged_data %>%
 filter(genus_species == "Kalanchoe_luciae") %>%
 group_by(PCB_number) %>%
 mutate(percent = percent_area/sum(percent_area)) %>% ungroup() %>%
 group_by(peak_ID) %>%
 summarize(percent_area = mean(percent)) %>% ungroup() %>%
 mutate(percent_area = percent_area/sum(percent_area)*100) -> temp

temp$Genus <- "Kalanchoe"
temp$Species <- "luciae"
temp$genus_species <- "Kalanchoe_luciae"
```

```
temp$PCB_number <- "NA"
chemical_bloom_merged_data %>%
  filter(genus_species != "Kalanchoe_luciae") -> chemical_bloom_merged_data
chemical_bloom_merged_data <- rbind(chemical_bloom_merged_data, temp)


## Mean Kalanchoe_thrysiflora
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_12",]
chemical_bloom_merged_data %>%
  filter(genus_species == "Kalanchoe_thrysiflora") %>%
  group_by(PCB_number) %>%
  mutate(percent = percent_area/sum(percent_area)) %>% ungroup() %>%
  group_by(peak_ID) %>%
  summarize(percent_area = mean(percent)) %>% ungroup() %>%
  mutate(percent_area = percent_area/sum(percent_area)*100) -> temp

temp$Genus <- "Kalanchoe"
temp$Species <- "thrysiflora"
temp$genus_species <- "Kalanchoe_thrysiflora"
temp$PCB_number <- "NA"
chemical_bloom_merged_data %>%
  filter(genus_species != "Kalanchoe_thrysiflora") -> chemical_bloom_merged_data
chemical_bloom_merged_data <- rbind(chemical_bloom_merged_data, temp)


## Mean Opuntia_sp
chemical_bloom_merged_data %>%
  filter(genus_species == "Opuntia_sp") %>%
  group_by(PCB_number) %>%
  mutate(percent = percent_area/sum(percent_area)) %>% ungroup() %>%
  group_by(peak_ID) %>%
  summarize(percent_area = mean(percent)) %>% ungroup() %>%
  mutate(percent_area = percent_area/sum(percent_area)*100) -> temp

temp$Genus <- "Opuntia"
temp$Species <- "sp"
temp$genus_species <- "Opuntia_sp"
temp$PCB_number <- "NA"
chemical_bloom_merged_data %>%
  filter(genus_species != "Opuntia_sp") -> chemical_bloom_merged_data
chemical_bloom_merged_data <- rbind(chemical_bloom_merged_data, temp)


## Mean Hosta_sp
```

```
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_85",]
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_92",]
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_93",]
chemical_bloom_merged_data %>%
  filter(genus_species == "Hosta_sp") %>%
  group_by(PCB_number) %>%
  mutate(percent = percent_area/sum(percent_area)) %>% ungroup() %>%
  group_by(peak_ID) %>%
  summarize(percent_area = mean(percent)) %>% ungroup() %>%
  mutate(percent_area = percent_area/sum(percent_area)*100) -> temp

temp$Genus <- "Hosta"
temp$Species <- "sp"
temp$genus_species <- "Hosta_sp"
temp$PCB_number <- "NA"
chemical_bloom_merged_data %>%
  filter(genus_species != "Hosta_sp") -> chemical_bloom_merged_data
chemical_bloom_merged_data <- rbind(chemical_bloom_merged_data, temp)

## Filter Graptopetalum_paraguayense
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_31",]
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_187",]

## Filter Brassica_oleracea
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_16",]
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_35",]
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_95",]

## Filter Ananas_comosus
```

```
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_5",]
chemical_bloom_merged_data <-
chemical_bloom_merged_data[!chemical_bloom_merged_data$PCB_number ==
"PCB_34",]

### Let's check on the number of unique species at this point (82 as of 2/16/2022):
unique(chemical_bloom_merged_data$genus_species)


##############################################################################
##############################################################################
##############################################################################
##############################################################################
##############################################################################



## Build a tree of the genus_species in your dataset (note that 10 species are dropped cuz
they're not in the tree)
genus_tree <- fortify(buildTree(
  scaffold_type = "newick",
  scaffold_in_path = "http://thebustalab.github.io/data/angiosperms.newick",
  members = unique(chemical_bloom_merged_data$genus_species)
))
chemical_bloom_merged_data_with_tree <- full_join(genus_tree,
chemical_bloom_merged_data, by = c("label" = "genus_species"))

## Let's check the number of species again (as of 2/16/22 it's 72, which makes sense, it's
82-10)
chemical_bloom_merged_data_with_tree %>% filter(isTip == TRUE) %>% select(label)
%>% unique()

## Draw phylogenetic tree
phylo_tree <- ggtree(chemical_bloom_merged_data_with_tree) +
  geom_tiplab(
          aes(label = label), size = 1.5, align = TRUE,
          hjust = 1, offset = 90, geom = "label", label.size = 0
  ) +
  coord_cartesian(xlim = c(0,450))+
  scale_y_continuous(expand = c(0,0.5)) +
  theme_void() +
  theme(plot.margin = unit(c(0.6,0,6,0), "cm"))
phylo_tree
```

```
## Make the order of genus in "chemical_bloom_merged_data_with_tree" match the order of the tree
genus_species_in_this_tree <- filter(chemical_bloom_merged_data_with_tree, isTip == TRUE)$label
genus_species_in_this_tree_ordered <- unique(genus_species_in_this_tree[order(filter(chemical_bloom_merged_data_with_tree, isTip == TRUE)$y, decreasing = TRUE)])

chemical_bloom_merged_data_with_tree$label <- factor(chemical_bloom_merged_data_with_tree$label, levels = rev(genus_species_in_this_tree_ordered))
# chemical_bloom_merged_data_with_tree <- filter(chemical_bloom_merged_data_with_tree, label != "<NA>")
# occurrences$genus_species

chemical_bloom_merged_data_with_tree$compound_class <- "Other compounds"


chemical_bloom_merged_data_with_tree$compound_class[chemical_bloom_merged_data_with_tree$peak_ID %in% c(
  "beta-sitosterol",
  "glutinol",
  "friedelin",
  "beta-amyrin acetate",
  "germanicol acetate",
  "lupeol acetate",
  "lupeol",
  "germanicol",
  "beta-amyrin",
  "beta-amyrone",
  "alpha-amyrin acetate",
  "triterpenoid acetate",
  "alpha-amyrin",
  "alpha-amyrone",
  "oleanolic acid",
  "ursolic acid",
  "unknown triterpenoid",
  "unknown triterpenoid acetate"
)] <- "Triterpenoid"

chemical_bloom_merged_data_with_tree$compound_class[chemical_bloom_merged_data_with_tree$peak_ID %in% c(
  "26-hexacosanal",
  "28-octacosanal",
  "29-nonacosanal",
  "30-triacontanal",
```

```
  "31-hentriacontanal",
  "32-dotriacontanal",
  "34-tetratriacontanal"
)] <- "Aldehydes"

chemical_bloom_merged_data_with_tree$compound_class[chemical_bloom_merged_dat
a_with_tree$peak_ID %in% c(
  "19-nonadecane",
  "21-heneicosane",
  "24-tetracosane",
  "25-pentacosane",
  "26-hexacosane",
  "27-heptacosane",
  "28-octacosane",
  "29-nonacosane",
  "30-triacontane",
  "31-hentriacontane",
  "32-dotriacontane",
  "33-tritriacontane",
  "34-tetratriacontane",
  "35-pentatriacontane",
  "37-heptatriacontane",
  "39-nonatriacontane",
  "40-tetracontane",
  "41-hentetracontane"
)] <- "Alkanes"

chemical_bloom_merged_data_with_tree$compound_class[chemical_bloom_merged_dat
a_with_tree$peak_ID %in% c(
  "24-tetracosan-1-ol",
  "25-pentacosan-1-ol",
  "26-hexacosan-1-ol",
  "27-heptacosan-1-ol",
  "28-octacosan-1-ol",
  "29-nonacosan-1-ol",
  "29-nonacosan-10-ol",
  "30-triacontan-1-ol",
  "31-hentriacontan-9-ol",
  "32-dotriacontan-1-ol",
  "34-tetratriacontan-1-ol",
  "40-tetracontan-1-ol"
)] <- "Alcohols"

chemical_bloom_merged_data_with_tree$compound_class[chemical_bloom_merged_dat
a_with_tree$peak_ID %in% c(
  "16-palmitic acid",
```

```
  "18-stearic acid",
  "20-arachidic acid",
  "22-behenic acid",
  "24-lignoceric acid",
  "26-cerotic acid",
  "28-montanic acid",
  "30-melissic acid",
  "32-lacceroic acid",
  "36-hexatriacontanoic acid"
)] <- "Fatty acids"

chemical_bloom_merged_data_with_tree$compound_class[chemical_bloom_merged_dat
a_with_tree$peak_ID %in% c(
  "16-hentriacontan-1-one",
  "26-hexacosan-1-one",
  "28-octacosan-1-one",
  "29-nonacosan-11-one",
  "31-hentriacontan-12-one",
  "31-hentriacontan-16-one",
  "diketone?"
)] <- "Ketones"

#chemical_bloom_merged_data_with_tree <-
chemical_bloom_merged_data_with_tree[!chemical_bloom_merged_data_with_tree$co
mpound_class == "Other compounds",]
#chemical_bloom_merged_data_with_tree <-
chemical_bloom_merged_data_with_tree[!chemical_bloom_merged_data_with_tree$co
mpound_class == "Aldehydes",]
#chemical_bloom_merged_data_with_tree <-
chemical_bloom_merged_data_with_tree[!chemical_bloom_merged_data_with_tree$co
mpound_class == "Ketones",]
###
chemical_bloom_merged_data_with_tree$compound_class[chemical_bloom_merged_dat
a_with_tree$peak_ID %in% c(
#  "plastic contaminant",
#  "beta-sitosterol/30-triacontan-1-ol",
#  "30-triacontatetraenoic acid",
#  "5847_dioctyl phthalate",
#  "unknown",
#  "sugar",
#  "disaccharaide (sucrose)",
#  "contaminant",
#  "3076_dis(2-ethyl hexyl)isophthalate",
#  "eicosyl eicosanoate(20-20)/octadecyl docosanoate (18-22)",
#  "20628_octadecanoic acid-2,3-dihydroxypropylester",
  ####
```

```
#  "3076-di(2-ethyl hexyl)isophthalate",
#  "glycerol monostearate"

#)] <- "Others"

## Draw a heat map of the compound occurrences
#heat_map <- ggplot(data = occurrences) +
#         geom_tile(aes(x = compound_common_name, y = genus, fill =
compound_relative_abundance)) +
#  theme(axis.text.x = element_text(angle = 90, hjust = 1, size =5))
#heat_map

## Set color scheme
# colors <- c("red", "pink", "grey")
# names(colors) <- c("major", "minor", "none detected")

## Draw a heat map of the compound occurrences
#occurrences <- occurrences[!is.na(occurrences$genus),]
#occurrences <- occurrences[!is.na(occurrences$compound_common_name),]
# yline_positions <- as.numeric(occurrences$genus) + 0.5
# xline_positions <- as.numeric(factor(occurrences$compound_common_name)) + 0.5
chemical_bloom_merged_data_with_tree <-
chemical_bloom_merged_data_with_tree[!is.na(chemical_bloom_merged_data_with_tre
e$label),]
chemical_bloom_merged_data_with_tree <-
chemical_bloom_merged_data_with_tree[!is.na(chemical_bloom_merged_data_with_tre
e$peak_ID),]
heat_map <- ggplot(data = chemical_bloom_merged_data_with_tree) +
  geom_tile(
          data = chemical_bloom_merged_data_with_tree,
          aes(x = peak_ID, y = label, fill = percent_area),
          color = "black"
  ) +
  #geom_hline(aes(yintercept = yline_positions), color = "black", size = 0.05) +
  #geom_vline(aes(xintercept = xline_positions), color = "black", size = 0.05) +
  scale_fill_gradient(low = "grey90", high = "black") +
  scale_x_discrete(name = "Compounds") +
  scale_y_discrete(name = "") +
  theme_classic() +
  facet_grid(.~compound_class, scales = "free_x", space = "free_x") +
  theme(
          axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5,size = 5),
          axis.text.y = element_blank(),
          # plot.margin = unit(c(0,0,0,-0.55), "cm")
          plot.margin = unit(c(0,0,-3,-0.55), "cm")
          # panel.grid.minor = element_line(color = "black", size = 1)
```

```
  )
heat_map

## Draw a bar plots triterpenoid occurrences for each species

## Make the order of species in "right_plot" match the order of the tree
# occurrences$compound_class <-
just_compounds$compound_class[match(occurrences$compound_common_name,
just_compounds$compound_name)]
# occurrences$genus <- factor(occurrences$genus, levels =
rev(genus_in_this_tree_ordered))

## Make the "right_plot"
chemical_bloom_merged_data_with_tree %>%
  filter(label != "NA") %>%
  dplyr::select(label, peak_ID, compound_class) %>%
  unique() %>%
  ggplot() +
  geom_bar(aes(y = label, fill = compound_class))+
  scale_y_discrete(name = "Genus_species")+
  theme_bw()+
  theme(
        axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 5),
        axis.text.y = element_text(size = 5),
        plot.margin = unit(c(0.6,0,0,-0.5), "cm") ) -> right_plot
##panel.grid.minor = element_line(color = "black", size = 1)
right_plot

## Make the "top_plot"
chemical_bloom_merged_data_with_tree %>%
  filter(peak_ID != "NA") %>%
  dplyr::select(label, peak_ID, compound_class) %>%
  ggplot() +
  geom_bar(aes(x = peak_ID)) +
  scale_y_continuous(name = "")+
  theme_bw()+
  theme(
        axis.text.x = element_blank(),
        #element_text(angle = 90, hjust = 1),
        axis.text.y = element_blank(),
        # plot.margin = unit(c(0,0,-0.5,-0.4), "cm")
        plot.margin = unit(c(0,0,-0.5,-0.4), "cm")
        #panel.grid.minor = element_line(color = "black", size = 1)
  ) +
  facet_grid(.~compound_class, space = "free_x", scales = "free_x") -> top_plot
top_plot
```

```
blank <- ggplot() + theme_void()

#final_plot1 <- cowplot::plot_grid(
#phylo_genus_tree, blank, heat_map_genus, blank, right_plot_genus,
#align = "h", axis = "b",
#rel_widths = c(0.5, -0.215, 1.5,-0.23, 0.5),
#nrow = 1
#)
#final_plot1

#final_plot2 <- cowplot::plot_grid(
#NULL, NULL, top_plot_genus, NULL, NULL,
#final_plot1,
#rel_heights = c(0.2, 0.5),
#nrow = 2
#)
#final_plot2

#final_plot1 <- cowplot::plot_grid(
#cowplot::align_plots(NULL, NULL, top_plot_genus, NULL, NULL, align = "h"),
#cowplot::align_plots(phylo_genus_tree, blank, heat_map_genus, blank,
right_plot_genus, align = "h"),
#rel_widths = c(0.5, -0.215, 1.5,-0.23, 0.5),
#ncol = 5
#)
#final_plot1

## Draft 1:
# final_plot1 <- cowplot::plot_grid(
#   phylo_genus_tree, blank, heat_map_genus, blank, right_plot_genus,
#   align = "h", axis = "b",
#   rel_widths = c(0.5, -0.215, 1.5,-0.23, 0.5),
#   nrow = 1
# )

col1 <- cowplot::align_plots(blank, phylo_tree, align = "v", axis = "l")
col2 <- cowplot::align_plots(blank, blank, align = "v", axis = "l")
col3 <- cowplot::align_plots(top_plot, heat_map, align = "v", axis = "r")
col3 <- cowplot::align_plots(col3[[1]], col3[[2]], align = "v", axis = "l")
col4 <- cowplot::align_plots(blank, blank, align = "v", axis = "l")
col5 <- cowplot::align_plots(blank, right_plot, align = "v", axis = "l")

bottom_row <- cowplot::align_plots(col1[[2]], col2[[2]], col3[[2]], col4[[2]], col5[[2]],
align = "h", axis = "b")
```

```
final_plot1 <- cowplot::plot_grid(
  col1[[1]], col2[[1]], col3[[1]], col4[[1]], col5[[1]],
  bottom_row[[1]], bottom_row[[2]], bottom_row[[3]], bottom_row[[4]],
bottom_row[[5]],
  # align = "h", axis = "b",
  rel_widths = c(0.5, -0.03, 1.5,-0.1, 0.4, 0.5, -0.03, 1.5,-0.1, 0.4),
  nrow = 2,
  rel_heights = c(1,4)
)

ggsave(
  filename = "/Users/dienngyen/Desktop/test4.png", plot = final_plot1,
  device = "png", width = 22, height = 12, units = "in", dpi = 600
)
```