Branch mutation rates and selection give insights on gene family trees: N-methyltransferases, FAD2 enzymes and terpenes

Daniella Santos Muñoz

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree of
Master of Computer Science Specialization in Bioinformatics

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Abstract

Gene trees enable the detailed study of the evolution of genes. These trees can be additive, based purely on mutational counts, or ultrametric, where the branches represent the time elapsed. The purpose of this thesis will be to relate evolutionary rate changes to gene function by examining the effects of a tree transformation on gene trees. The study will focus on the coffee genomes *C. arabica, C. eugenioides* and *C. canephora.* Three gene family trees will be studied: N-methyltransferases, FAD2 enzymes and terpenes. An additive tree can be transformed into an ultrametric tree using a transformation method. By taking each additive tree branch length and dividing the branch by the ultrametric branch length we can get a branch mutation rate. To quantify functional divergence, we compared aligned sequences of genes from groups of interest to calculate the ratio of non-synonymous mutations to synonymous mutations (Kn/Ks). Kn/Ks was found to correlate with the branch mutation rate in some but not all groups of interest. In groups in which correlation was present, there was also a correlation between Kn/Ks and branch length. The branch mutation rate can be used in some cases to further prove that functional divergence is present.

Acknowledgements

I would like to thank my supervisor David Sankoff for the years of mentorship he provided me through undergraduate and graduate school. Thank you for expanding my knowledge of bioinformatics, and for the opportunities you provided me to learn the various areas of bioinformatics through conferences and workshops.

Thank you to my mother and family in supporting me through my studies. You always supported my quest of learning and my growth as a person, no matter how hard it got.

I would also like to thank Jarkko T. Salojarvi, Victor A. Albert, and the other researchers in the Arabica Coffee Genome Consortium for including me in their project, and for providing data I could use for my thesis.

Contents

List of Figures

List of Tables

List of Abbreviations

CoGe	Comparative Genomics research platform
FAD	Fatty acid oleate desaturase
Kn (dN)	Non-synonymous mutations
Ks (dS)	Synonymous mutations
PAML	Phylogenetic analysis by maximum likelihood
PASTA	Practical Alignment using Saté and TrAnsitivity
PL	Penalized likelihood
MYA	Million years ago
NCBI	National Center for Biotechnology Information
NMT	N-methyltransferase

Chapter 1

Introduction

1.1 Background

Gene trees enable the detailed study of the evolution of genes. These trees can be additive, based purely on mutational counts, or ultrametric, where branches repesent the time elapsed. To study evolutionary rates, an additive tree must be transformed into an ultrametric tree through the use of a transformation method. One such method we will use is Penalized Likelihood (PL). The transformation causes branches to shorten or lengthen to reach the time requirement of ultrametric trees. The change in branch length gives an indication of the evolutionary rate. We will apply Penalized Likelihood to gene family trees to visualize which genes have accelerated or decelerated evolutionary rates, allowing us to relate these rates to gene functions.

Previously, I studied a method in Santos Muñoz, Lam, and Sankoff (2019) that can infer phylogenies from the coding sequence of genomes while at the same time dating these phylogenies, the method was found to be particularly useful for working with genomes that had undergone whole genome duplication events. The method involved a transformative method, Penalized Likelihood (Sanderson, 2002), transforming the initial additive tree into an ultrametric tree. Additive trees can be based on mutations but may not always meet the ultrametric time requirement. The transformation maintained the correct topology of the additive tree while meeting the ultrametric time requirement, allowing for the phylogeny to be dated and for the relationship between topology and branch length change to be studied (Klopfstein, Massingham, and Goldman, 2017). The change in branch length, from the additive tree to the ultrametric tree, gave an indication of the evolutionary rate of the species (external branches). A significant increase in branch length can indicate acceleration of the evolutionary rate, and vice versa. By applying the transformation on to a gene tree (relating to a function), the evolutionary rate changes can be related to gene function.

A phylogeny based on mutations has branches that correspond to the number of mutations. The more mutations are present, the longer the branch. So more closely related genes (i.e. less mutations between them) will have a shorter tree distance between them. While more diverging branches will have a longer branch distance between them.

By applying a transformation to the branch lengths we can relate the branches to a time constraint, relating the amount of mutations present in the branch to a time scale. By taking the original branch length (pre-transformation) and dividing the branch by the transformed branch length (post-transformation) we can get a mutation rate (number of mutations / time) for the branch. This branch mutation rate is what we will study.

The mutation rate can vary at many levels. At the genome level, genes can vary in mutation rate as some genes can be under more selection pressure than other genes. At the species population level, different populations can have different mutation rates depending on the selection pressure the population faces and the amount of genetic drift present (Lynch et al., 2016).

Two types of nucleotide mutations we will be looking at are synonymous (Ks or dS) and non-synonymous (Kn or dN) mutations. Non-synonymous mutations are a change in the nucleotide sequence that causes a change in the amino acid it encodes, while synonymous mutations do not cause changes in the amino acids. A change in the amino acids leads to changes in the resulting protein, which can cause a change in function of the protein.

When natural selection is applied to these changed proteins it sometimes works to optimize the protein structure and function towards what is being selected for. Since the force of selection on mutations can change depending on the drivers of selection, the pattern of Kn mutations can be complicated and erratic (Ohta, 1994).

As synonymous mutations do not cause a change in proteins, Ks mutations are less affected by selection than Kn mutations are. However weak selection can affect both Kn and Ks mutations. Selection acts on synonymous mutations by optimizing codon usage and GC nucleotide content (Ohta, 1994). Synonymous mutations can accumulate over time (Reis and Yang, 2013), as such they can be used for calculating evolutionary distance between lineages (Wolf et al., 2009). While Ks mutations are able to accumulate freely, Kn mutations are constrained by the types of amino acids that can take place within a protein due to the protein's structure and function. This causes the number of Kn mutations to be smaller than the number of Ks mutations (Reis and Yang, 2013).

By taking the ratio of non-synonymous mutations to synonymous mutations (Kn/Ks) functional divergence can be quantified. If the number of non-synonymous mutations is greater than the number of synonymous mutations ($Kn/Ks > 1$) then there is an overall change in function, indicating positive selection. If the number of non-synonymous mutations is less than the number of synonymous mutations ($Kn/Ks < 1$) then amino acid changes are constrained, signifying purifying selection (Wolf et al., 2009).

1.2 Purpose and Objectives

The purpose of this thesis will be to relate evolutionary rate changes to gene function by examining the effects of a tree transformation on gene trees. The gene trees that will be studied are composed of genes of similar function. The branch changes caused by the transformation of an additive gene tree to an ultrametric tree will be compared to Kn/Ks substitution rates. We will study if the comparison gives an indication of the relation of evolutionary rate to tree transformation branch changes. Branches with a smaller branch mutation rate (branch length increased by the transformation) will have a smaller Kn/Ks ratio, while branches with a larger branch mutation rate (branch length decreased by the transformation) have a larger Kn/Ks. A larger branch mutation rate signifies more mutations over less time. If more mutations have occurred, the Kn/Ks of a sequence will be larger.

Branches that will be compared will be determined through the selection of groups of branches of interest. These groups will be mainly composed of shorter branches (meeting a lower threshold length) or composed of longer branches (meeting a higher threshold length). To test Kn/Ks, we want to make sure at least one gene (terminal branch) in the group does not meet the threshold to see a trend between the genes that meet the threshold and those that do not meet the threshold.

Dating a gene tree was explored in this study by applying a known divergence time point to the ultrametric tree.

In our results, Kn/Ks was found to correlate to the branch mutation rate in some but not all groups of interest. In groups in which correlation was present, there was also a correlation between Kn/Ks and branch length.

Chapter 2

Application

We'll be applying our study of trees to plant genomes. The genomes of interest to us were produced by our collaborators in the Arabica Coffee Genome Consortium and pertain to the *Coffea* genus: *Coffea arabica, Coffea canephora* and *Coffea eugenoides.*

Coffea species are the first step in the coffee production, with each species having its own unique qualities and taste. A component of coffee that makes it so desirable, along with its taste, is caffeine. Caffeine is a stimulant that, at low doses, can increase attention and concentration, but at high doses can cause anxiety and neurosis (Bolton and Null, 1981). Coffee beans are not the only source of caffeine. Other sources of caffeine include tea leaves (*C. sinensis)*, and cocoa beans (*T. cacao).* Even though these species pertain to different parts of the world (coffee from Africa, tea from Asia and cacao from South America) they all produce the same caffeine molecule. However, different biosynthetic pathways are involved, leading to convergent evolution being a possible explanation (Huang et al., 2016). Possible reasons as to why caffeine may have appeared independently are due to the Chemical Defense Theory and Allelopathic Theory. By the Chemical Defense Theory, high concentrations of caffeine protect soft tissues in the plant from predators. By the Allelopathic Theory, the presence of caffeine on the coats of seeds stops the germination of other seeds (Ashihara and Crozier, 2001). Important to the biosynthetic pathway of caffeine production in coffee is N-methyltransferases (Denoeud et al., 2014). Multiple N-methyltransferases (NMT) are involved in the final steps of the biosynthetic pathway of caffeine, catalyzing the methylation of xanthosine, then methylation of 7mX and finally the methylation of theobromine. Theobromine is then methylated to become caffeine (Ogawa et al., 2001; Denoeud et al., 2014). The components of coffee that add to its taste and aroma are aromatic oils, these are composed mainly of terpenes, and fatty acids. Linoleic acid is a fatty acid in the coffee bean that is produced by the oleate desaturase FAD2 enzyme. In the coffee plant, FAD2 is linked to six genes, while in *A. thaliana*, which has a common ancestor with coffee, FAD2 is linked to one gene. The increase in FAD2 genes is caused by multiple tandem duplications (Denoeud et al., 2014). Genes related to N-methyltransferase, to the oleate desaturase FAD2 enzyme, and to terpene synthase were of focus in the thesis.

Chapter 3

Materials and Methods

3.1 Data

We used genomes from 23 plant species (Table 3.1), focusing on the *Coffea* species. Genomes were collected from NCBI, CoGe, Phytozome, Dryad, individual sources and provided to us by our collaborators from the Arabica Coffee Genome Consortium (De Kochko, Crouzillat, and Consortium, 2015). Data provided by the Arabica Coffee Genome Consortium included trees for each of the gene families and aligned gene sequences used in the production of these trees. Gene families of genes related to N-methyltransferase, to the oleate desaturase FAD2 enzyme, and to terpene synthase were of focus in the project.

3.2 Pipeline

Starting from an additive tree, the tree is transformed into an ultrametric tree (Section 3.4). The additive and ultrametric trees are then compared (Section 3.5) to produce a table containing branch information and comparisons. The branch comparisons were used for statistical testing of the comparison of each species' genes against the rest of the genes in the tree (Section 3.9), and for comparing Kn/Ks against the branch changes in groups of

TABLE 3.1: Plant species used in study (n = 23). References (ref.) for each genome are provided.

Family	Species	Common Name	Ref.
Amborellaceae	*Amborella trichopoda*	Amborella	Albert et al., 2013
Ranunculaceae	*Aquilegia coerulea*	Columbine	Filiault et al., 2018
Brassicaceae	*Arabidopsis thaliana*	Thale cress	Lamesch et al., 2011
Rubiaceae	*Coffea arabica*	Eugenioides	Denoeud et al., 2014
Rubiaceae	*Coffea eugenioides*	Arabica	Denoeud et al., 2014
Rubiaceae	*Coffea canephora*	Robusta	Denoeud et al., 2014
Salicaceae	*Populus trichocarpa*	Poplar	Du et al., 2015
Rosaceae	*Prunus persica*	Peach	Verde et al., 2013
Solanaceae	*Solanum lycopersicum*	Tomato	Sato et al., 2012
Malvaceae	*Theobroma cacao*	Cacao	Motamayor et al., 2013
Lentibulariaceae	*Utricularia gibba*	Bladderwort	Lan et al., 2017
Vitaceae	*Vitis vinifera*	Grape	Jaillon et al., 2007
Nelumbonaceae	*Nelumbo nucifera*	Sacred lotus	Ming et al., 2013
Oleaceae	*Olea europaea*	Olive tree	Unver et al., 2017
Convolvulaceae	*Ipomea nil*	Japanese morning glory	Hoshino et al., 2016
Pedaliaceae	*Sesamum indicum*	Sesame	Wang et al., 2016
Euphorbiaceae	*Ricinus communis*	Castor bean	Rivarola et al., 2011
Apocynaceae	*Catharantus roseus*	Madagaskar periwinkle	Kellner et al., 2015
Apocynaceae	*Calotropis gigantea*	Giant calotrope	Hoopes et al., 2018
Gelsemiaceae	*Gelsemium sempervirens*	Yellow yessamine	Franke et al., 2019
Nyssaceae	*Camptotheca acuminata*	Happy tree	Zhao et al., 2017
Actinidiaceae	*Actinidia chinensis*	Kiwifruit	Huang et al., 2013
Theaceae	*Camellia sinensis*	Tea	Xia et al., 2019

interest in each tree. Groups of interest to the study were derived from the additive tree in Section 3.6, and were mainly composed of short branches or long branches. The Kn/Ks was calculated for each branch in each group of interest (Section 3.7.2) and then compared against the branch lengths and branch mutation rates for each branch in each group of interest (Section 3.7.3). From the ultrametric tree and a known branching point the tree can be dated (Section 3.8). Implementation was done in R with methods from the Ape (Paradis, Claude, and Strimmer, 2004), Phylobase (Bolker et al., 2011) and Phangorn (Schliep, 2011) packages.

3.3 Gene Tree Reconstruction

Our collaborators at the Arabica Coffee Genome Consortium produced the gene trees for the three gene families through the subsequent procedure. Gene trees, based on functions of interest, were built by using amino acid sequences of the plant species used in the project. A domain search of the sequences identified all orthogroups that had hits to each of the functions of interest. The union of these orthogroups, for each function, was passed through PASTA (Mirarab et al., 2015), a multiple sequence alignment tool. Default settings were used for PASTA. The alignments, measured in substitutions per site, were used as input to FastTree (Price, Dehal, and Arkin, 2010), a tool that infers phylogenetic trees from a maximum likelihood approximation. Three unrooted additive trees were produced in total: a tree containing genes (n = 699) involved in NMT function, a tree containing genes (n = 1440) involved in terpene synthase function, and a tree containing genes (n = 652) involved in FAD function.

3.4 Phylogenetic Transformation

To transform an additive tree into an ultrametric tree, we used Penalized Likelihood. PL functions by applying a smoothing parameter and rate penalty to an additive tree to add a time constraint to the tree, making the tree ultrametric. Before applying the transformation, the tree had to be prepared first. To apply the transformation to a tree, there cannot be any branches with a length of zero. Since some of the trees had some branches with a length of zero a constant was added to all branches of each tree. The same constant, 0.1, was added to all trees. Figures showing the original tree structure do not have this constant added, but all data used from the tree includes this added constant.

Cross-validation was then performed on the tree to find the optimal smoothing parameter, λ. Cross-validation successively drops branches when testing λ values. The λ values tested were 1×10^{-1}, 1, 1×10^{1}, 1×10^{2}, ..., 1×10^{6}. A sum of the D2 attributes (influence of each observation on the date estimates) was taken for each cross-validation fold. The absolute value of the median sum was used as the λ for PL.

Since the cross-validation process is computationally intensive (Sanderson, 2002), trees as large as the gene function trees used in this project can take a long time to run. On an iMac with a 2.3GHz dual-core Intel Core i5 and 8GB ram, the NMT and FAD tree transformations took approximately 16 hours, and the terpene tree took approximately 6 days to complete.

3.5 Tree Comparisons

The branch lengths from the transformed tree were compared and tabulated with the respective branch lengths of the original tree, and the difference and quotient (branch mutation rate) of the branch lengths was taken. Differences are transformed branch length minus original branch length, as seen in Equation 3.1. Branch mutation rate is calculated as the original branch length divided by the transformed branch length, as seen in Equation 3.2. If the value of the quotient is greater than 1 then the transformed branch is shorter, so there is branch shortening. If the value is 1 then no change has occurred. If the value is less than 1 then the transformed branch is longer, so there is branch lengthening.

$$difference = original\ branch - transformed\ branch \tag{3.1}$$

$$branch\ mutation\ rate = \frac{original\ branch}{transformed\ branch} \tag{3.2}$$

Before comparing branch lengths, the original tree branch lengths were scaled to be within the range of the transformed tree branch lengths. An ultrametric tree always has a distance of 1 from each tip to the root, since each tip is equidistant from the root; however, additive trees do not need to meet this requirement so the distance from each tip to the root can vary by tree. Along with the branch change information, the location of the branch in the phylogeny and the name of the branch were saved. If branches were external, they were named by the gene ended at; otherwise, they were called internal.

3.6 Gene Groups of Interest

Groups of interest were selected for each tree to study the Kn/Ks of the branches of these groups. Tree groups were determined by grouping tree tips (genes) with the shortest lengths or longest lengths of the original tree.

Branch length thresholds for groupings take into account the 0.1 constant that was added to all branches of the original tree, but not the original branch length scaling that was applied to the data.

Groups with a lower terminal branch threshold had branches shorter than 0.12 (the shortest possible length is 0.1). This value allowed for groupings to form (enough genes adjacent to form a group) but not create too many groups to heavily outnumber the longer length groups. Groups with an upper (or high) terminal branch threshold had branches longer than 0.4. This value allowed for groupings to form (enough genes adjacent to form a group because branches with longer lengths are more scattered throughout the tree).

The branches are skewed towards shorter lengths (distribution), so the shorter length groups only incorporate the shortest branches, while the longer length groups have more variability in the branch length. To have a similar number of groups for each type of branch length in each tree, grouping criteria (branch length threshold and number of genes in group) were adjusted. Lower threshold groups have branch lengths between 0.12 and 0.1, while upper threshold groups have most branches longer than 0.4. The goal was to have an almost equal number of groups between the two types of groups for each tree (the number of total groups between trees could vary however) so that there would be an even amount of data between the two groups.

Other criteria that groups had to meet were: contain at least 4 genes that meet the branch length threshold, and more than half of the group meets the threshold. Genes that were noted of interest in these groups were genes that did not meet the threshold and coffee genes. The tree group selection algorithm (Algorithm 1) was used to identify the groups based on threshold number, threshold type (lower or upper), and minimum and average number of genes in the group that meet threshold.

For each group of interest of each tree, each gene had its terminal branch (tip) branch mutation rate B and gene-to-node branch mutation rate B_{ng} value calculated. The gene-to-node branch mutation rate value is the quotient of the sum from the tip (gene) g in a group to the parent node n of the group (the originator of the group). This value is calculated by, first, adding the original branch lengths o of all the branches from the parent node to the tip and by, second, adding the transformed branch lengths t of all the branches from the parent node to the tip, and lastly by taking the quotient of these two sums. For each group, the mean and median branch mutation rates and gene-to-node values were calculated. The gene-to-node branch mutation rate formula is shown in Equation 3.3. The gene-to-node value gives the cumulative branch mutation rate from a gene to its group parent node.

$$B_{ng} = \frac{\sum_{i=n}^{g} o_i}{\sum_{i=n}^{g} t_i} \tag{3.3}$$

Algorithm 1: Tree group selection. Groups are subtrees of the tree, in which groups contain terminal branches (tips) that must meet a length requirement. If the length requirement type is "lower", group tip branches must be shorter than a threshold value given. If the length requirement type is "upper", group tip branches must be longer than a threshold value given. A minimum number and percentage of tip branches must meet the requirement. The Descendant and Ancestor functions were imported through the Phangorn package.

```
Input :
        T  additive tree
        B  branch comparison table
        value  threshold value terminal branch lengths must meet
        type  threshold type ("upper" or "lower")
        min  minimum number of terminal branches in group that must meet
        threshold
        p  percentage of terminal branches in group that must meet threshold
Output:
        G  list of groups (groups defined by the root node of the group)
$G \leftarrow \emptyset$                                        ▷ initialize empty set of groups
for each node in T do
    $desc \leftarrow Descendants(T, node, type = "tips")$       ▷ descendant tips of node
    $count \leftarrow 0$                                        ▷ number of tips that meet threshold
    for each d in desc do
                                                                ▷ branch length of descendant d
        $branch \leftarrow table["Right.node" = d, "Original.branch.length"]$
                                                                ▷ check threshold
        if (type = "lower" and branch < value) or (type = "upper" and branch > value)
         then
            count++
        end
    end
                                                                ▷ check if group of node is valid
    if $count \geq min$ and $count > num(desc) * p$ then
        $G \leftarrow add(node)$
    end
end
                                                                ▷ keep most ancestral nodes for each group
for each node in G do
    $anc \leftarrow Ancestors(T, node, type = "all")$            ▷ all ancestral nodes of node
    if anc in G then
        $G \leftarrow remove(node)$
    end
end
return G
```

3.7 Kn/Ks

3.7.1 Protein to Codon

To calculate Kn/Ks values for the branches, nucleotide sequences were needed for each gene, however we only had amino acid sequence data. These amino acid sequences were used to query the genome (from which the sequence derives) to find the matching nucleotide sequence. The search was performed by uploading the genomes to CoGe and querying the sequences with CoGeBlast. By using tblastn (protein sequence query) in CoGeBlast, the amino acid sequences could be reverse translated to DNA. Query parameters used were: an e-value of 1e-5, a word size of 10, a gap cost of 12 existence and 1 extension, and queries were filtered. BLOSSUM62 with composition-based statistics was used. Stop codons were not included. If results could not be found through CoGEBlast, Blast was used directly from NCBI. The amino acid sequences were queried through tblastn with the default genome for the species queried. To check that the nucleotide sequences did in fact match the amino acid sequences, the nucleotide sequences were translated and compared to the data amino acid sequences.

Ambiguous amino acids in the sequences (denoted by 'X') were replaced with the corresponding codons found in the nucleotide sequence resulting from the query. Some amino acid sequences were ancestral and did not pertain to any one species, therefore a corresponding nucleotide sequence could not be found for these amino acid sequences. Groups of interest that contained an ancestral sequence could not be used in the Kn/Ks comparison experiments.

Resulting nucleotide sequences were matched to the aligned amino acid sequences by adding gaps ('-') at the appropriate locations. For each single gap present in the aligned amino acid sequence, three gaps were added to the nucleotide sequence so that codons would be aligned.

3.7.2 Kn/Ks calculation

Paml (Yang, 1997) was used to calculate the ratio of synonymous to nonsynonymous mutations, Kn/Ks. The alignments (nucleotide version) along with the group subtrees were used in CodeML to calculate the ratio. CodeML options selected were: seqtype = 1 (codon), clock = 0 (no clock present) and model = branch model.

Branch models allow the Kn/Ks ratio to vary among branches of the phylogeny. Different branch models can be applied depending on how branches in the tree are selected. These models can be used for detecting positive selection on particular lineages in the tree. The free-ratio model (model = 1 on CodeML) assumes that each branch has a different rate. To apply this model, no particular branches are selected. This model was used for all groups. The two-ratio (model = 2 on CodeML) model assumes that selected branches have a rate different from the branches not selected. To apply this model, all selected branches were denoted with "#1". This model was used for all groups that had highlighted branches. The three-ratio (model = 2 on CodeML) model assumes that selected branches have a rate different from the branches not selected, and that selected branches can have a rate different from each other. To apply this model, all selected branches were denoted with a different "#n" (where n=1,2,3,...). This model was used for all groups that had highlighted branches.

3.7.3 Kn/Ks Branch Comparison

Paml results were compared in R against group branch comparison data. Kn/Ks values of each branch were plotted against branch mutation rate values and branch lengths of each branch. This was done for each ratio model. Both all branches in the group and only tip branches were each plotted. t-test R correlation values and p-values were calculated for each comparison.

3.8 Dating

A tree can be dated by using a known divergence time point and scaling the tree with respect to that time. The known time of a divergence point between two groups of species can be found in literature (derived from fossil timings). A group of species can consist of only one species or several species (e.g. family, order, etc). The divergence point is found in the tree by finding the most recent common ancestor between the two groups of species, i.e. the ancestor node contains both groups in its descendants (Algorithm 2). For a species tree, this method is much more straightforward because there is only one occurrence of each species in the tree; therefore, there is only one divergence point between two groups of species in a species tree. However, gene trees are much less straightforward. Gene trees can contain several genes per species, a species is represented several times in a gene tree. To find the divergence time point in a gene tree, first, all nodes are considered that contain a gene pertaining to each group in their descendants. After these nodes are found, only the most recent nodes are considered, since ancestral nodes to the recent nodes would be redundant. Since the recent nodes can have various distances from the tips of the tree (time zero), the distance of the divergence point is calculated by aggregating the values through an average, as seen in Algorithm 2 or by taking the maximum distance.

Trees were dated by using the split between the *C. roseus (*Apocynaceae family) and coffee species (Rubiaceae family). The average date of 79 million years was used, taken from the Time Tree of Life (Kumar et al., 2017). The maximum (i.e. the first occurrence of divergence) and the average of the distance to the tips of the most recent nodes ancestral

Algorithm 2: Gene tree dating. Tree is dated based on a divergence point, which is determined from two given groups. Returned divergence point distance is an average aggregate. Descendant and Children functions were imported through the Phangorn package.

```
Input :
        T    ultrametric tree
        B    branch comparison table
        G1   group 1 (list of Strings)
        G2   group 2 (list of Strings)
Output:
        dist   distance from tips (time 0)
$G1.genes, G2.genes \leftarrow \emptyset$        ▷ initialize empty set of group gene nodes
                                                  ▷ go through table and find gene nodes
for i to nrow(table) do
    for each g in G1 do
                                                  ▷ check that tip label contains string g
        if g in table[i,"Node.label"] then
            $G1.genes \leftarrow add(table[i,"Node"])$
        end
    end
    for each g in G2 do
        if g in table[i,"Node.label"] then
            $G2.genes \leftarrow add(table[i,"Node"])$
        end
    end
end
                    ▷ Find internal nodes that have at least one descendant from each group
$common \leftarrow \emptyset$
for each node in T do
    $desc \leftarrow Descendants(T, node, type = "tips")$        ▷ Descendant tips of node
          ▷ check that an intersection exists between the node's descendants and each group
    if $\exists(desc \cap G1.genes)$ and $\exists(desc \cap G2.genes)$ then
        $common \leftarrow add(node)$
    end
end
            ▷ Only keep least ancestral nodes (to have most recent divergence point)
$d.point \leftarrow \emptyset$ for each node in common do
    $desc \leftarrow Descendants(T, node, type = "all")$         ▷ All descendants of node
          ▷ check that an intersection does not exist for the node's descendants and common
    if $\exists!(desc \cap common)$ then
        $d.point \leftarrow add(node)$
    end
end
                      ▷ Find distance for each divergence point (common node)
$d.point.dist \leftarrow \emptyset$ for each node in d.point do
    $dist \leftarrow 0$                                         ▷ initialize distance
    $kids \leftarrow Children(T, node)$                           ▷ Both child nodes of node
    $current \leftarrow node$                                     ▷ start pointer at node
                                          ▷ Calculate distance from node to tree tip
    while $\exists kids$ do
        $dist \leftarrow$dist$+table["left.node" = currentand"right.node" = kids[1], "branch.length"]$
        $current \leftarrow kids[1]$      ▷ move pointer to a child node (both kids have same distance)
        $kids \leftarrow Children(T, current)$                ▷ Both child nodes of current node
    end
    $d.point.dist \leftarrow add(dist)$
end
return mean(d.point.dist)
```

to both species groups was used to find the dating point in the tree. The dating point was then used to scale the time scale to the tree.

3.9 Branch Change Statistical Testing

Statistical tests were applied to the branch change table of the gene trees to detect significant branch length differences of the genes, species, and groups of species. One-sample t-tests were applied to the branch change table to detect a significant difference (two-sided test), shortening ('less' alternative), or lengthening ('greater' alternative) for each branch. The test was applied to the original branch change table containing both internal and external branches, and to a branch change table only containing external branches. The table containing all branches allowed for all branches to be involved in the comparison. The table containing only external branches allowed for only the external gene branches to be compared. To detect if the genes of a species shortened or lengthened significantly as a whole, the genes of a species were compared against the rest of the genes in the tree. Welch two-sample t-tests were applied to the external branch changes table to detect significant branch lengthening ('greater' alternative) or shortening ('less' alternative) of species genes in comparison to the rest of the genes in the tree. The same test was done for comparing groups of species against the rest of the tree.

Statistical tests were implemented in R, and the Broom package (Robinson, 2014) was used to tabulate the results of the statistical tests. Statistical tests were done for the difference in branch length (Equation 3.1) and the quotient of the branch length (Equation 3.2).

Chapter 4

Results

In this chapter we present the results of our methodology from Chapter 3, in which our goal is relating evolutionary rate to gene function by studying the effects of applying a transformation to gene trees.

4.1 Trees

Trees (Figures 4.1,4.2, and 4.3) are presented with their additive branch lengths and coloured according to the log of the branch mutation rate. The additive tree before transformation is also referred to as the original tree. The resulting ultrametric tree after transformation is also referred to as the transformed tree. For branch mutation rates (Equation 3.2), if the transformed branch length was greater than the original branch length then the branch mutation rate was between zero and one; while a transformed branch length less than the original branch length could cause the branch mutation rate to be much larger than one. A log transformation was performed on the branch mutation rate to subdue this skew. Purple to red branches represent negative branch mutation rates, while blue to teal branches represent positive branch mutation rates in the tree figures. Groups of interest were highlighted in the tree and numbered in order of appearance starting at the top of

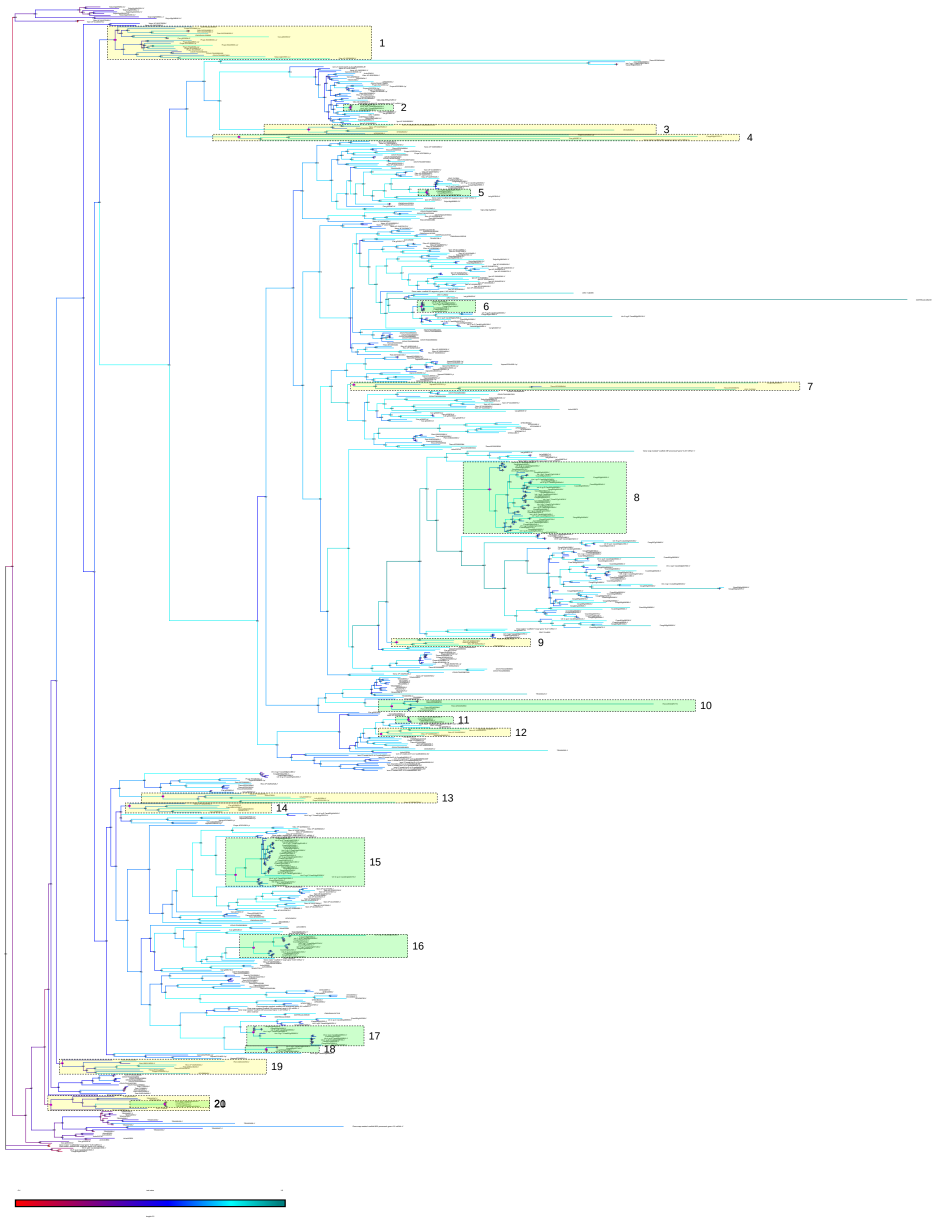

FIGURE 4.1: NMT additive tree with branch mutation rates and groups of interest. Colour scale based on log branch mutation rates. Groups highlighted in boxes and threshold denoted by yellow (meet upper threshold) or green (meet lower threshold). Groups numbered from top to bottom and group parent node highlighted in a magenta circle (n=689). Details can be seen by zooming in.

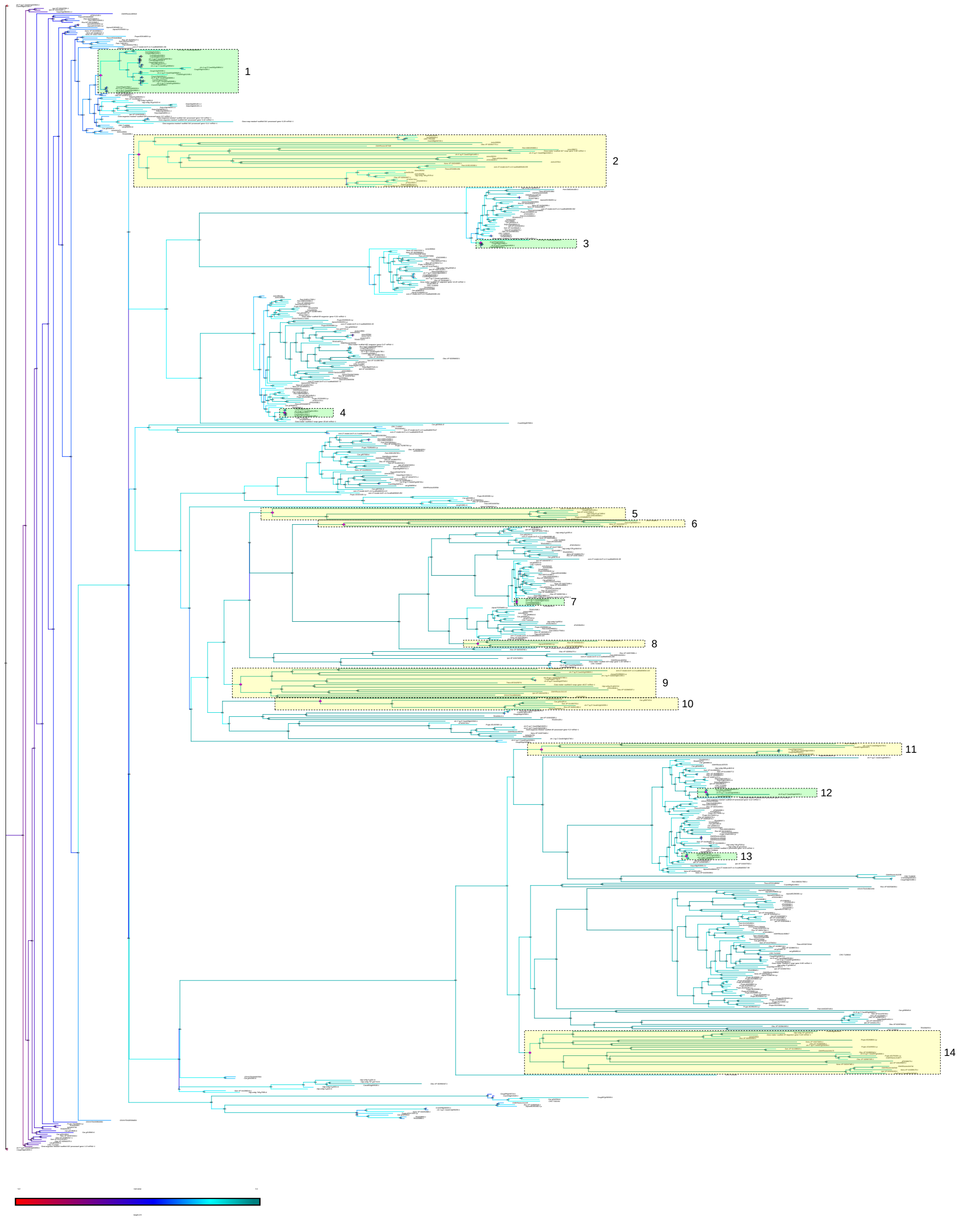

FIGURE 4.2: FAD additive tree with branch mutation rates and groups of interest. Colour scale based on log branch mutation rates. Groups highlighted in boxes and threshold denoted by yellow (meet upper threshold) or green (meet lower threshold). Groups numbered from top to bottom and group parent node highlighted in a magenta circle (n=658). Details can be seen by zooming in.

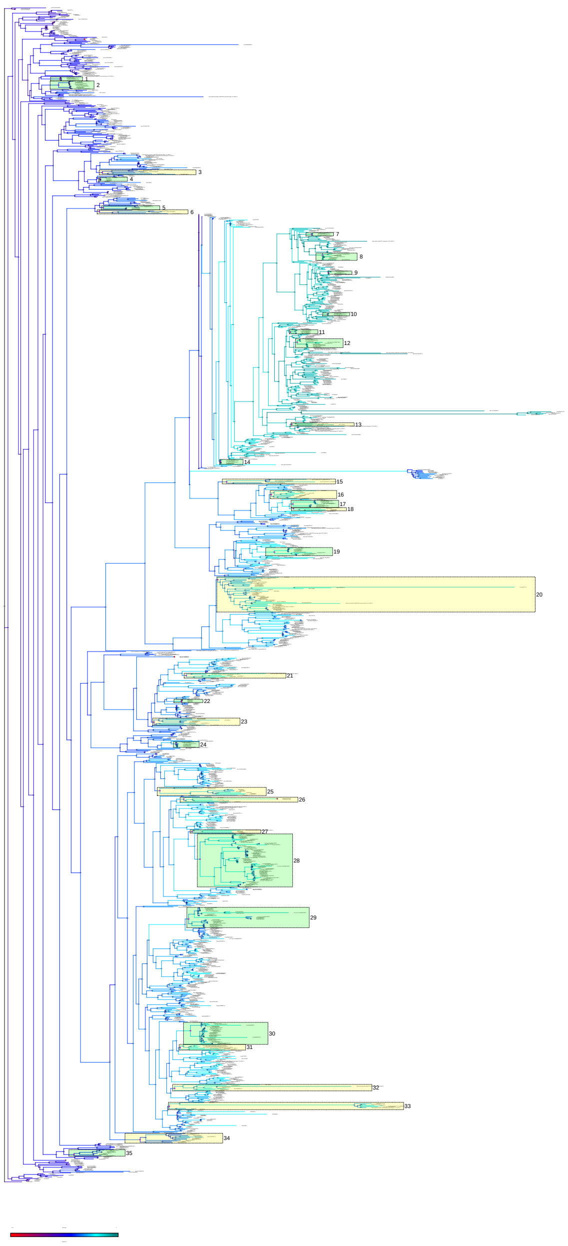

FIGURE 4.3: Terpene additive tree with branch mutation rates and groups of interest. Colour scale based on log branch mutation rates. Groups highlighted in boxes and threshold denoted by yellow (meet upper threshold) or green (meet lower threshold). Groups numbered from top to bottom and group parent node highlighted in a magenta circle (n=1440). Details can be seen by zooming in.

the tree. Groups that met a lower threshold were coloured green, while groups that met a higher threshold were coloured yellow.

The NMT tree had the highest number of genes in its groups of interest, 209 out of a total 689 genes, followed by the Terpene tree, 373 out of a total 1440 genes, and finally the FAD tree had 151 out of a total 658 genes. The NMT tree had the highest ratio of coffee genes to total genes (220 out of 689 genes), followed by the Terpene tree (308 out of 1440 genes) and then the FAD tree (109 out of 658 genes). However, the Terpene tree had the highest ratio of coffee genes in groups of interest out of the total number of coffee genes (202 out 308 genes), followed by the FAD and NMT trees which had similar ratios, 65 out of 109 genes and 131 out of 220 genes, respectively. The groups are further examined in the next section.

4.2 Gene Groups of Interest

In order to test the Kn/Ks values against the branch mutation rate values, we selected groups of interest in each tree, as per Section 3.6. Each group for each tree are presented in Appendix A. The tip branch mutation rate and gene-to-node branch mutation rate values can be found for each gene within the group, along with the mean and median aggregates of these values for the group.

In the groups of each tree (Figures 4.4, 4.5 and 4.6), branch mutation rate and gene-to-node values of low threshold groups were overall lower than the values of high threshold groups. This trend was most apparent in the NMT and FAD tree groups. The trend could be seen for most groups in the Terpene tree but not all. In low threshold groups, there was a greater difference between branch mutation rate values and gene-to-node values for each group than in high threshold groups. When a greater difference was present between branch mutation rate values and gene-to-node values for a group, the gene-to-node value tended to be higher than the branch mutation rate value. Low threshold groups were almost all composed of only coffee genes, and groups containing coffee genes only consisted of coffee genes, while high threshold groups mostly contained genes of species other than coffee. Coffee genes present in high threshold groups tended to be lower than the threshold. Ancestral genes (genes not pertaining to a particular species) were only found in high threshold groups.

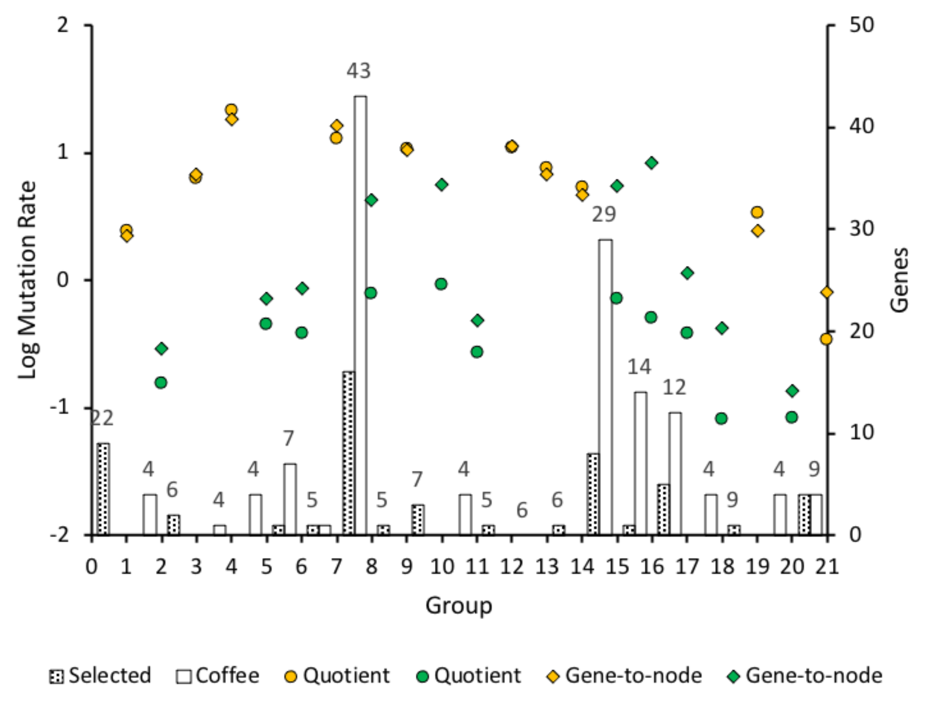

FIGURE 4.4: NMT groups of interest with mean tip branch mutation rate values and gene-to-node values. Group threshold denoted by yellow (meet upper threshold) or green (meet lower threshold). Tip branch mutation rate values (quotient) denoted by circles and gene-to-node values denoted by diamonds. The log of each branch mutation rate and gene-to-node value was calculated before taking the mean of each group. The number of genes selected (genes that do not meet the group threshold) are denoted by shaded bars and coffee genes (*Coffea* genus) are denoted by white bars to the left of the group label. Total group gene number is shown above bars (total number of genes in these groups is 209).

In the low threshold groups of the NMT tree (Figure 4.4), all groups had branch mutation rates lower than zero. Conversely, high threshold groups had branch mutation rates higher than zero; however, group 21 had a branch mutation rate and gene-to-node value lower than zero, which could be due to group 21 containing the low threshold group 20.

In the FAD tree groups (Figure 4.5), all green groups had lower values than the values of the yellow groups, however only two groups (4 and 7) had log quotient values below zero.

The Terpene tree groups (Figure 4.6) were not as defined between the two types of

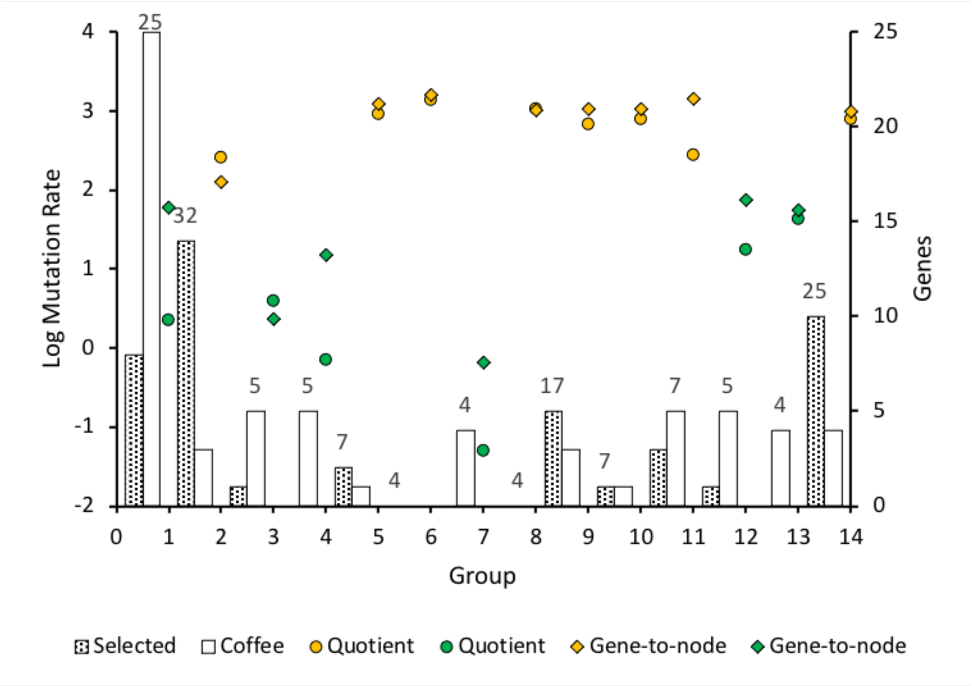

FIGURE 4.5: FAD groups of interest with mean tip branch mutation rate values and gene-to-node values. Group threshold denoted by yellow (meet upper threshold) or green (meet lower threshold). Tip branch mutation rate values (quotient) denoted by circles and gene-to-node values denoted by diamonds. The log of each branch mutation rate and gene-to-node value was calculated before taking the mean of each group. The number of genes selected (genes that do not meet the group threshold) are denoted by shaded bars and coffee genes (*Coffea* genus) are denoted by white bars to the left of the group label. Total group gene number is shown above bars (total number of genes in these groups is 151).

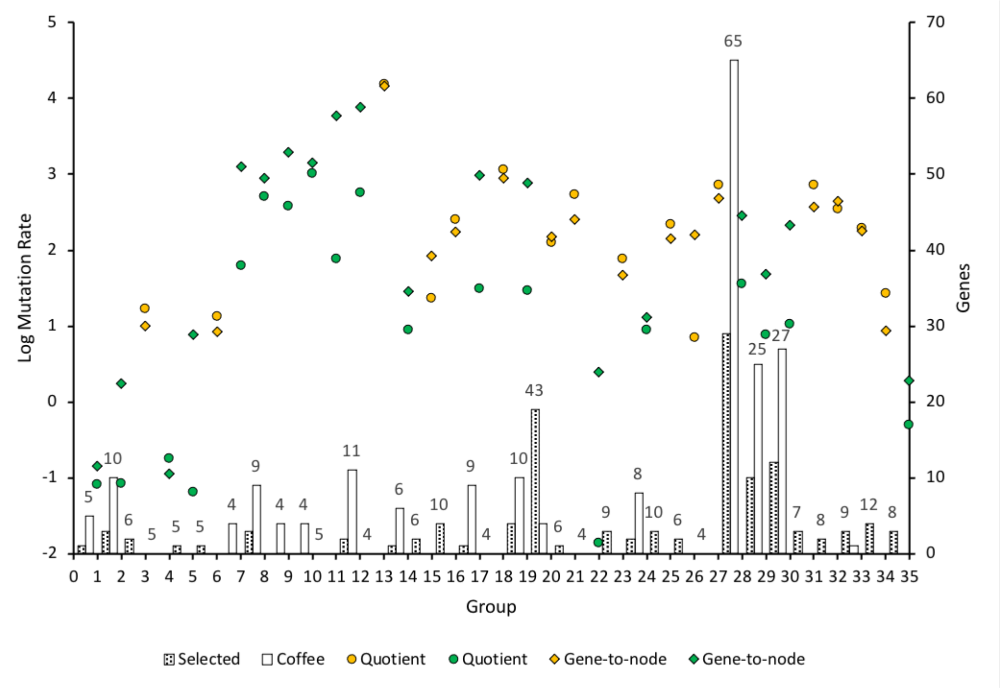

FIGURE 4.6: Terpene groups of interest with mean tip branch mutation rate values and gene-to-node values. Group threshold denoted by yellow (meet upper threshold) or green (meet lower threshold). Tip branch mutation rate values (quotient) denoted by circles and gene-to-node values denoted by diamonds. The log of each branch mutation rate and gene-to-node value was calculated before taking the mean of each group. The number of genes selected (genes that do not meet the group threshold) are denoted by shaded bars and coffee genes (*Coffea* genus) are denoted by white bars to the left of the group label. Total group gene number is shown above bars (total number of genes in these groups is 373).

groups branch mutation rate wise when compared to the groups in the NMT and FAD trees. Many low threshold groups had similar values to high threshold groups. Most Terpene groups had branch mutation rate values greater than zero, except for groups 1, 2, 4, 5, 22 and 35.

4.3 Comparison of Kn/Ks and Branch Mutation Rate

Comparisons of Kn/Ks and branch mutation rate were performed for each tree (Tables 4.1, 4.2 and 4.3). Groups that were included in the comparison either contained coffee genes, had one or multiple genes not meeting the group threshold, and did not contain any ancestral genes. Many low threshold groups in each tree met all three of these criteria. Few high threshold groups met all three criteria; therefore, these groups did not have to contain coffee genes to be considered. For each group, the branch Kn/Ks was compared to the length of the branch and to to the branch quotient value.

When testing the three branch models, free-, two- and three-ratio models, groups that did not contain genes that did not meet the group threshold could only be tested with the free-ratio model. The two- and three-ratio models required branches to be selected, and our selection criteria was genes that did not meet the group threshold. Groups that were only tested with the free-ratio model for this reason were NMT groups 2, 5, 11, 18 and 20, FAD groups 1, 4, 7 and 13, and Terpene groups 7, 9 and 10.

For choosing the high threshold groups to include, NMT groups 7, 14 and 19 and Terpene groups 6, 23 and 25 could not be included because a matching nucleotide sequence for the amino acid sequence could not be found for *T. cacao* in neither CoGeBlast nor NCBI Blast. Of the only two eligible high threshold groups of the FAD tree, only group 10 could be used because in group 5 a matching nucleotide sequence for the amino acid sequence could not be found for *U. gibba.* For all ratio models, the Terpene groups 21, 28, 31 and 32 did not produce Kn/Ks results in Paml. For the free-ratio model, NMT group 8 did not produce Kn/Ks results in Paml.

There was a correlation present between Kn/Ks and branch length and between Kn/Ks and branch mutation rate in all trees, but not in every group. Some groups had a positive correlation while others had a negative correlation.

In the Kn/Ks comparison of the NMT groups (Table 4.1), group 6 was the only group that showed correlation for all ratio models. In Figure 4.7, it can be seen that the longest external branch has the highest Kn/Ks of the external branches. The high correlation for group 6 can be explained by the longest branch, which is the only branch that does not meet the group threshold, having such a large Kn/Ks value. Other groups either had correlation present in the two- and three-ratio models but not the free-ratio model, or the group could only be tested for the free-ratio model and therefore had no data for the other ratio models. Groups 17 and 21 were the only groups that only had correlation present in the two-ratio model. Most groups for the NMT tree had a positive correlation, however groups 12 and 17 had a negative correlation. The three groups with the highest correlation, 6, 12 and 16, all had only one gene each that was selected (not meeting group threshold). The groups that did not have a correlation present, had many selected genes within their group, except for groups 3 and 9 which had two and one selected genes each,respectively.

The FAD tree (Table 4.2) had the fewest number of groups of interest, and had few

TABLE 4.1: NMT Kn/Ks (dN/dS) tip branch comparison. For each branch, its Kn/Ks and branch mutation rate (BMR) or branch length were compared for the free-, three- and two-ratio branch models. R correlation values and p-values from the linear regression were calculated for each group. R values >= 0.7 and <= - 0.7 were highlighted. Low threshold groups are highlighted green and high threshold groups are highlighted yellow. An α value of 0.05 was used to detect significance.

Group	**Free-ratio**				**Three-ratio**				**Two-ratio**			
	dN/dS vs branch		**dN/dS vs BMR**		**dN/dS vs branch**		**dN/dS vs BMR**		**dN/dS vs branch**		**dN/dS vs BMR**	
	R	**p-value**	**R**	**p-value**	**R**	**p-value**	**R**	**p-value**	**R**	**p-value**	**R**	**p-value**
2	**1.00**	2.2E-16	**1.00**	2.3E-04	/	/	/	/	/	/	/	/
5	0.25	7.5E-01	0.36	6.4E-01	/	/	/	/	/	/	/	/
6	**0.99**	2.5E-05	**0.89**	6.9E-03	**0.99**	6.0E-06	**0.76**	4.9E-02	**0.99**	6.0E-06	**0.76**	4.9E-02
8	/	/	/	/	0.04	8.0E-01	0.14	3.8E-01	0.59	2.7E-05	**0.82**	1.2E-11
11	**1.00**	2.6E-04	**0.99**	1.3E-02	/	/	/	/	/	/	/	/
15	-0.14	4.6E-01	-0.08	6.7E-01	**0.73**	5.8E-06	0.57	1.2E-03	0.46	1.2E-02	0.66	1.1E-04
16	-0.14	6.4E-01	0.06	8.3E-01	**1.00**	2.2E-16	**0.81**	4.0E-04	**1.00**	2.2E-16	**0.81**	4.0E-04
17	-0.13	6.8E-01	0.13	6.9E-01	0.38	2.3E-01	0.05	8.9E-01	**-0.80**	1.9E-03	**-0.75**	5.2E-03
18	**1.00**	2.5E-06	**1.00**	1.8E-03	/	/	/	/	/	/	/	/
20	-0.30	7.0E-01	-0.08	9.2E-01	/	/	/	/	/	/	/	/
3	0.57	2.4E-01	0.50	3.1E-01	0.44	3.9E-01	0.65	1.6E-01	0.44	3.8E-01	0.63	1.8E-01
9	**0.73**	1.6E-01	0.57	3.2E-01	0.51	3.8E-01	0.46	4.4E-01	0.51	3.8E-01	0.46	4.4E-01
12	-0.06	9.2E-01	0.19	7.6E-01	**-0.81**	9.5E-02	**-0.77**	1.2E-01	**-0.81**	9.5E-02	**-0.77**	1.2E-01
21	-0.39	2.9E-01	-0.37	3.3E-01	-0.39	2.9E-01	-0.37	3.3E-01	**0.98**	1.6E-06	**0.87**	2.5E-03

possible groups for Kn/Ks comparisons. Only groups 3 and 12 had correlation present, but not with the free-ratio model. Group 3 had some correlation present between Kn/Ks and branch length but not between Kn/Ks and branch mutation rate. Groups 3 and 12 had one selected gene each but group 3 had a positive correlation present while group 12 had a negative correlation present.

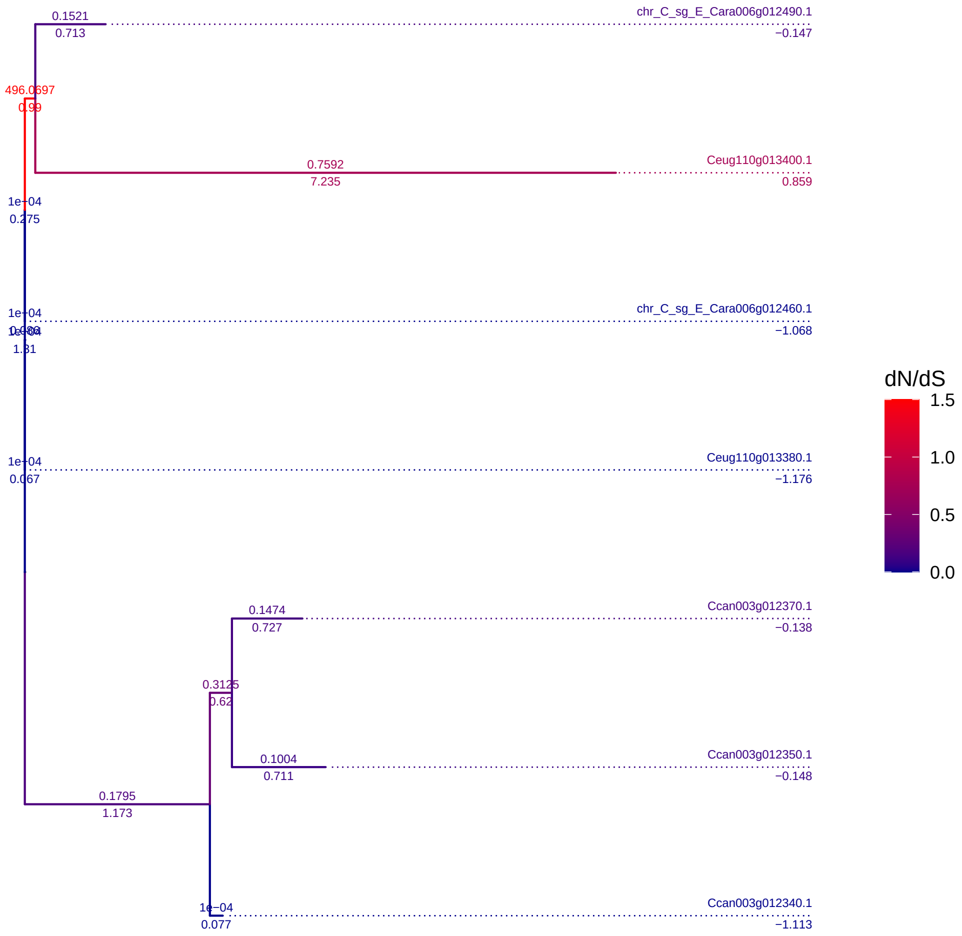

FIGURE 4.7: Kn/Ks NMT group 6 subtree. Kn/Ks is denoted by dN/dS. Log tip branch mutation rate values are shown under the tip gene label. For internal branches, the value above the branch is dN/dS and the value below is the branch mutation rate. Branches are coloured according to the dN/dS value.

TABLE 4.2: FAD Kn/Ks (dN/dS) tip branch comparison. For each branch, its Kn/Ks and branch mutation rate (BMR) or branch length were compared for the free-, three- and two-ratio branch models. R correlation values and p-values from the linear regression were calculated for each group. R values >= 0.7 and <= - 0.7 were highlighted. Low threshold groups are highlighted green and high threshold groups are highlighted yellow. An α value of 0.05 was used to detect significance.

Group	**Free-ratio**				**Three-ratio**				**Two-ratio**			
	dN/dS vs branch		**dN/dS vs BMR**		**dN/dS vs branch**		**dN/dS vs BMR**		**dN/dS vs branch**		**dN/dS vs BMR**	
	R	**p-value**	**R**	**p-value**	**R**	**p-value**	**R**	**p-value**	**R**	**p-value**	**R**	**p-value**
1	-0.18	3.80E-01	0.08	6.90E-01	/	/	/	/	/	/	/	/
3	**0.79**	1.1E-01	0.46	4.40E-01	**1.00**	3.10E-06	**0.70**	1.90E-01	**1.00**	3.10E-06	**0.70**	1.90E-01
4	-0.40	5.00E-01	-0.54	3.50E-01	/	/	/	/	/	/	/	/
7	0.00	/	0.00	/	/	/	/	/	/	/	/	/
12	-0.28	6.50E-01	0.16	8.00E-01	**-1.00**	1.50E-06	**-0.82**	8.80E-02	**-1.00**	1.50E-06	**-0.82**	8.80E-02
13	0.26	7.40E-01	0.19	8.10E-01	/	/	/	/	/	/	/	/
10	-0.22	6.30E-01	-0.53	2.20E-01	0.32	4.80E-01	0.40	3.70E-01	0.32	4.80E-01	0.40	3.70E-01

There were many Kn/Ks comparisons with a correlation present for the Terpene tree (Table 4.3). Although, only one group, group 2, had a correlation present for all ratio models. Groups 1, 12, 15 and 17 had a correlation present in the two and three-ratio models but not in the free-ratio model. Interestingly, for both the two- and three-ratio models, group 14 had a correlation present only in the Kn/Ks and branch length comparison, while group 19 only had a correlation present in the Kn/Ks and branch mutation rate comparison. Correlation for the high threshold groups was negative while only one group in the low threshold groups was negative (group 1). Contrary to the NMT and FAD tree Kn/Ks comparisons, the Terpene group with correlation present among all ratio models, group 2, had three selected genes in its groups, while other groups with correlation present had less selected genes in their groups.

Across the groups of each tree in which correlation was present in the Kn/Ks comparisons, the correlation of Kn/Ks against the branch length tended to be higher than the correlation of Kn/Ks against the branch mutation rate. The direction (positive or negative) of the correlation was the same as the correlation that was seen when comparing the Kn/Ks to the branch length. For groups that had a positive correlation of Kn/Ks to

TABLE 4.3: Terpene Kn/Ks (dN/dS) tip branch comparison. For each branch, its Kn/Ks and branch mutation rate (BMR) or branch length were compared for the free-, three- and two-ratio branch models. R correlation values and p-values from the linear regression were calculated for each group. R values >= 0.7 and <= - 0.7 were highlighted. Low threshold groups are highlighted green and high threshold groups are highlighted yellow. An α value of 0.05 was used to detect significance.

Group	**Free-ratio**				**Three-ratio**				**Two-ratio**			
	dN/dS vs branch		**dN/dS vs BMR**		**dN/dS vs branch**		**dN/dS vs BMR**		**dN/dS vs branch**		**dN/dS vs BMR**	
	R	**p-value**	**R**	**p-value**	**R**	**p-value**	**R**	**p-value**	**R**	**p-value**	**R**	**p-value**
1	-0.28	6.50E-01	-0.49	4.00E-01	**-0.98**	4.40E-03	**-0.98**	2.60E-03	**-0.98**	4.40E-03	**-0.98**	2.60E-03
2	**0.7**	2.50E-02	**0.81**	4.50E-03	**0.92**	1.80E-04	0.57	8.60E-02	**0.82**	3.80E-03	**0.71**	2.20E-02
7	**1**	5.00E-05	**1**	5.70E-04	/	/	/	/	/	/	/	/
8	-0.11	7.80E-01	-0.14	7.20E-01	**0.84**	4.60E-03	**0.85**	3.90E-03	**0.85**	3.50E-03	**0.92**	4.60E-04
9	-0.29	7.10E-01	-0.024	9.80E-01	/	/	/	/	/	/	/	/
10	**-0.73**	2.70E-01	**-0.72**	2.80E-01	/	/	/	/	/	/	/	/
12	-0.19	5.70E-01	0.027	9.40E-01	**0.99**	1.10E-09	**0.73**	1.10E-02	**0.94**	1.60E-05	**0.75**	7.30E-03
14	-0.0055	9.90E-01	0.29	5.80E-01	**0.98**	6.40E-04	0.68	1.40E-01	**0.98**	6.40E-04	0.68	1.40E-01
17	-0.18	6.50E-01	-0.18	6.40E-01	**0.98**	5.60E-06	**0.86**	3.00E-03	**0.98**	5.60E-06	**0.86**	3.00E-03
19	-0.17	6.40E-01	-0.11	7.50E-01	0.65	4.20E-02	**0.73**	1.70E-02	0.6	6.70E-02	**0.93**	8.00E-05
24	0.65	1.10E-01	0.63	1.30E-01	0.54	2.10E-01	0.24	6.00E-01	**0.74**	5.60E-02	0.3	5.10E-01
29	-0.1	6.20E-01	0.16	4.30E-01	0.0045	9.80E-01	0.31	1.30E-01	0.35	8.30E-02	**0.7**	1.10E-04
30	-0.09	6.60E-01	0.083	6.80E-01	0.13	5.20E-01	0.34	8.50E-02	0.41	3.40E-02	**0.74**	9.40E-06
35	-0.21	6.20E-01	-0.21	6.20E-01	**0.71**	5.00E-02	0.64	8.50E-02	0.67	6.90E-02	**0.82**	1.30E-02
3	-0.24	6.50E-01	-0.27	6.00E-01	-0.38	4.60E-01	-0.3	5.60E-01	**-0.75**	8.40E-02	**-0.89**	1.70E-02
15	-0.61	2.00E-01	-0.19	7.20E-01	**-0.92**	9.00E-03	**-0.82**	4.70E-02	**-0.92**	9.00E-03	**-0.82**	4.70E-02
16	-0.4	2.60E-01	-0.25	4.90E-01	-0.48	**1.60E-01**	-0.3	4.00E-01	**-0.81**	4.50E-03	**-0.72**	1.90E-02
34	-0.15	6.40E-01	0.29	3.70E-01	-0.14	**6.60E-01**	0.18	5.70E-01	**-0.62**	3.20E-02	0.18	5.70E-01

branch length, as the branch mutation rate increased, the Kn/Ks increased. Therefore, branches that decreased in length after the transformation (larger branch mutation rate) had a higher rate of non-synonymous mutations. For groups that had a negative correlation of Kn/Ks to branch length, as the branch mutation rate increased, the Kn/Ks decreased. In these groups, branches that decreased in length after the transformation (larger branch mutation rate) had a lower rate of non-synonymous mutations.

4.4 Species Genes Statistical Testing

To observe how the genes of different species were affected by the tree transformation, statistical analyses were performed on the branch differences (Table 4.4) and the branch mutation rate (Table 4.5). The gene terminal branches of each species were compared to the rest of the genes in the tree to detect if any species had a significantly higher or lower branch difference and branch mutation rate.

TABLE 4.4: Welch two-sample t-test of the branch change differences of genes of each species. The genes of a species are compared against the rest of the genes in the tree. Branch length differences are measured in substitutions per site. P-values shown for species branch differences being greater or less than the rest of the genes in the tree. An α value of 0.05 was used to detect significance.

Species	NMT		FAD		Trep	
	greater	less	greater	less	greater	less
A. trichopoda	0.802	0.198	0.868	0.132	0.988	**0.012**
A. coerulea	0.204	0.796	0.659	0.341	0.585	0.415
A. thaliana	0.919	0.081	0.704	0.296	1.000	**2.36E-04**
C. arabica	**1.54E-07**	1.000	0.834	0.166	**2.40E-06**	1.000
C. eugenioides	0.122	0.878	**9.81E-10**	1.000	**3.92E-11**	1.000
C. canephora	**1.52E-07**	1.000	0.256	0.744	**6.69E-23**	1.000
P. trichocarpa	0.089	0.911	0.511	0.489	**1.35E-05**	1.000
P. persica	0.764	0.236	0.837	0.163	**0.025**	0.975
S. lycopersicum	0.572	0.428	**1.49E-03**	0.999	0.999	**5.71E-04**
T. cacao	0.907	0.093	0.309	0.691	0.882	0.118
U. gibba	0.876	0.124	0.112	0.888	0.971	**0.029**
V. vinifera	0.095	0.905	0.737	0.263	**5.50E-03**	0.994
N. nucifera	0.789	0.211	0.736	0.264	0.882	0.118
O. europaea	0.276	0.724	0.909	0.091	0.346	0.654
I. nil	0.253	0.747	0.518	0.482	0.391	0.609
S. indicum	0.423	0.577	**1.22E-11**	1.000	0.436	0.564
R. communis	0.629	0.371	0.475	0.525	0.999	**1.20E-03**
C. roseus	0.952	0.048	0.466	0.534	1.000	**6.73E-05**
C. gigantea	0.982	**0.018**	**6.37E-10**	1.000	1.000	**6.31E-05**
G. sempervirens	0.849	0.151	0.225	0.775	0.994	**5.59E-03**
C. acuminata	0.806	0.194	0.544	0.456	0.743	0.257
A. chinensis	0.726	0.274	0.366	0.634	0.883	0.117
C. sinensis	0.771	0.229	0.471	0.529	0.104	0.896

Genes with a greater branch difference, compared to the rest of the genes, signified that the genes had greater branch lengthening, while genes with a greater branch mutation rate signified that the genes had a greater branch shortening (i.e. less branch lengthening). Therefore, these measures, branch difference and branch mutation rate, worked opposite to each other (i.e. a greater branch mutation rate inferred a lesser branch difference), but not always. These opposite measures could be seen for some of the species, such as: *C. arabica* and *C. gigantea* in the NMT and Terpene trees; *C. eugenioides* in the FAD and Terpene trees; *A. thaliana, C. roseus* and *G. sempervirens* in the Terpene tree; and *S. indicum* in the FAD tree.

TABLE 4.5: Welch two-sample t-test of the branch mutation rate of genes of each species. The genes of a species are compared against the rest of the genes in the tree. Branch mutation rates are measured in substitutions per site. P-values shown for species branch mutation rates being greater or less than the rest of the genes in the tree. An α value of 0.05 was used to detect significance.

Species	NMT		FAD		Trep	
	greater	less	greater	less	greater	less
A. trichopoda	0.972	**0.028**	0.195	0.805	1.000	**1.87E-09**
A. coerulea	0.580	0.420	0.309	0.691	0.987	**0.013**
A. thaliana	**0.016**	0.984	**4.74E-03**	0.995	**0.024**	0.976
C. arabica	1.000	**2.67E-04**	0.685	0.315	0.998	**2.48E-03**
C. eugenioides	0.873	0.127	1.000	**1.25E-12**	0.977	**0.023**
C. canephora	0.955	**0.045**	0.999	**8.74E-04**	0.991	**8.71E-03**
P. trichocarpa	0.998	**1.62E-03**	0.506	0.494	0.828	0.172
P. persica	0.680	0.320	**0.057**	0.943	0.314	0.686
S. lycopersicum	0.508	0.492	0.681	0.319	0.547	0.453
T. cacao	0.191	0.809	0.538	0.462	0.722	0.278
U. gibba	**0.049**	0.951	0.149	0.851	0.080	0.920
V. vinifera	0.731	0.269	0.296	0.704	0.198	0.802
N. nucifera	0.912	0.088	0.207	0.793	0.157	0.843
O. europaea	0.451	0.549	0.077	0.923	0.527	0.473
I. nil	**0.017**	0.983	0.095	0.905	0.866	0.134
S. indicum	**0.011**	0.989	0.972	**0.028**	0.996	**4.22E-03**
R. communis	**0.030**	0.970	0.233	0.767	0.064	0.936
C. roseus	**0.034**	0.966	0.248	0.752	**0.027**	0.973
C. gigantea	**2.59E-03**	0.997	0.982	**0.018**	0.315	0.685
G. sempervirens	0.193	0.807	0.768	0.232	**0.024**	0.976
C. acuminata	0.646	0.354	0.949	**0.051**	1.000	**6.86E-05**
A. chinensis	0.409	0.591	0.983	**0.017**	0.316	0.684
C. sinensis	0.697	0.303	0.419	0.581	0.885	0.115

For the NMT tree, there was only two coffee species with a greater branch difference and one other species with a lesser branch difference. However, there were more species with a greater and a lesser branch mutation rate. In the FAD tree, there were some species with a significantly greater difference and none with a lesser difference. However, for the branch mutation rate, there were some species with a significantly greater branch mutation rate and many more species with a significantly lesser branch mutation rate. Lastly, in the Terpene tree, there were more species with a significantly lesser comparison for both branch differences and branch mutation rate, although not necessarily for the same species. There were more species with a significantly greater branch difference than for the branch mutation rate.

4.5 Dating

An ultrametric tree can be dated by using a known time reference. The distance into the tree in which the time reference is found can be used to add a time scale. A time reference of 79 MYA of the divergence of the coffee species and periwinkle (*C. roseus)* was used.

Timing the gene family trees (Figures 4.8, 4.9 and 4.10) was found to not be chronologically relevant. The root times surpassed the billions of years, which did not make sense biologically, and differed greatly between the trees. Using the maximum divergence point distance was more chronologically relevant for all trees. The FAD tree had the farthest root time at 205785 MYA, when the divergence point was calculated using the mean, while the root time was a lot less, 29984 MYA, when the maximum divergence point was used. The next farthest root time was seen in the Terpene tree, 15355 MYA, for the mean divergence point, and 1206 MYA for the maximum divergence point (which was the lowest root time). The NMT tree also had maximum divergence time, 2235 MYA, lower than the mean divergence time, 6053 MYA, but also had the smallest difference between the two times.

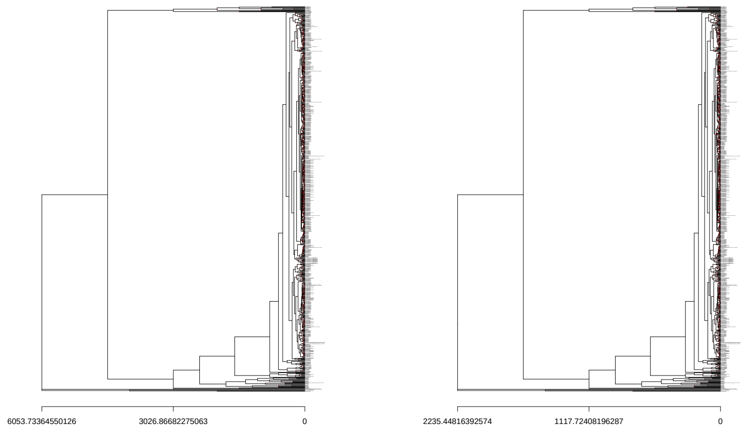

FIGURE 4.8: NMT dated ultrametric tree. A known time point of 79 MYA for the divergence between the coffee species and *C. roseus* was used to date the tree. Time scale in MYA. The time point was placed at the mean divergence distance (left) and maximum divergence distance (right).

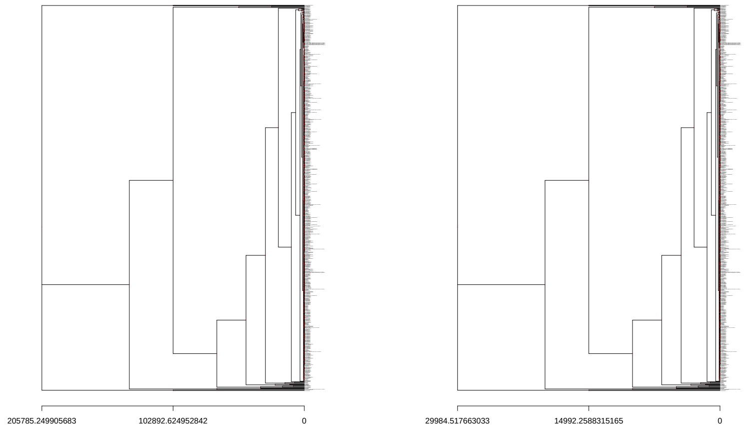

FIGURE 4.9: FAD dated ultrametric tree. A known time point of 79 MYA for the divergence between the coffee species and *C. roseus* was used to date the tree. Time scale in MYA. The time point was placed at the mean divergence distance (left) and maximum divergence distance (right).

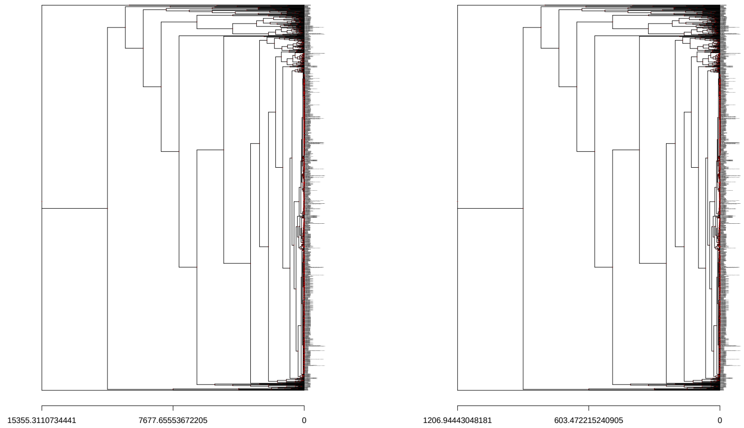

FIGURE 4.10: Terpene dated ultrametric tree. A known time point of 79 MYA for the divergence between the coffee species and *C. roseus* was used to date the tree. Time scale in MYA. The time point was placed at the mean divergence distance (left) and maximum divergence distance (right).

Chapter 5

Discussion

In this thesis, we studied the relation of evolutionary rate changes to gene function by examining how a tree transformation affected gene trees. Gene function divergence was quantified by comparing aligned gene sequences to calculate the ratio of non-synonymous mutations to synonymous mutations. To study the relation of evolutionary rate to gene function, we used the transformation of gene trees to calculate the branch mutation rate. Groups of interest were used to study the comparison of the branch mutation rate to Kn/Ks. For each branch in the group, the branch's branch mutation rate was compared to its Kn/Ks.

To study the relation of evolutionary rate changes to gene function in transformed trees, we used gene family trees based on plant genomes. The genomes of interest to us pertained to the *Coffea* genus: *Coffea arabica*, *Coffea canephora* and *Coffea eugenoides.* The gene functions we studied, of particular interest for coffee, were related to N-methyltransferase, to the oleate desaturase FAD2 enzyme, and to terpene synthase.

We found that Kn/Ks correlated with the branch mutation rate in some but not all groups of interest. In groups in which this correlation was present, there was correlation between Kn/Ks and branch lengths of the original tree. Similar results of correlation between Kn/Ks and branch length were seen in Wolf et al. (2009), in which Kn/Ks was found to be correlated with Kn but not Ks. The variation in correlation between Kn/Ks and the branch mutation across the groups of interest could be explained by the inconsistencies of Kn/Ks, which can vary across the tree (Yang, Nielsen, and Hasegawa, 1998).

By studying genes grouped by a function, there is a possibility of variation between the genes. Due to the number of genes included in the gene trees we studied and how each of these genes have evolved, there could be vast differences between the genes. To mitigate the gene variation, future studies could be done on gene trees that are more strict on the genes used, such as only using a limited number of genes and comparing its variation among species. Another way to mitigate variation would be to only compare genes of closely related species, such as studying a gene tree composed of only coffee species genes.

In this study, only gene functions important to the production of coffee were examined. Comparing the results for these functions with other gene trees based on conservative plant functions could be also used to examine how the correlation between Kn/Ks and branch mutation rate can vary among other functions. A systematic approach could also be done in which sequences are generated with varying degrees of divergence.

Limitations in our data affected the results. Due to only having amino acid data, and thus requiring the search of corresponding nucleotide sequences, some genes could not be studied as their nucleotide sequences could not be found. Ancestral genes were not designated to particular species, so could not be compared to a species' genome, and general searches matched multiple species but did not match the sequence completely. An exact match between amino acid sequence and nucleotide sequence was required to match the alignments given by our collaborators. The limitations caused by the ancestral genes greatly affected the study of the high threshold groups.

Due to the inconsistencies of the correlation between Kn/Ks and branch mutation rate, it was found that the branch mutation rate can be used in some cases to infer the presence of functional divergence. However, further proof will be needed to support the presence of functional divergence.

By using a known time reference point, we were able to add a time scale to the ultrametric trees. However, the resulting time scales of each tree were not biologically relevant as the root times surpassed billions of years. The time scales varied greatly between the trees as well. The dating of the ultrametric trees most likely did not work because the ultrametric trees were skewed towards a more recent divergence. Most divergence events were placed closer to current time (the tips of the tree), while the few ancestral divergence events were much farther (i.e. closer to the root). The distance between the events caused ancestral events to be dated much older than what is biologically possible, as land plants are believed to have first originated around 470 MYA (Lenton et al., 2016). There were also wide variations in time scales of the trees, which could be due to the evolution of each function. Each function, NMT involvement in caffeine production, FAD and Terpenes, most likely have different evolutionary histories as their purpose varies between each other. The pace of the evolutionary history of genes can affect the dating of the tree as the rate of molecular evolution can vary. The variation could be seen in Tables 4.5, where branch mutation rates varied by species and tree function. Another reason the dating may not have worked is due to the distribution of genes per species. For example, the trees had a larger number of coffee genes than if the number of genes were distributed equally by species; 16-32% of genes were coffee genes (percentage varied by tree), while the ratio of coffee species to total species in the trees is 13%.

Appendix A

Tree Groups of Interest

This appendix contains the data from which figures 4.4, 4.5 and 4.6 are derived. For each tree and group of interest, the table of tip of branch mutation rates (quotient) and gene-to-node quotients per gene is included. Lower threshold groups have a green header, while upper threshold groups have a yellow header. Genes that do not meet the group threshold have purple font and coffee genes are highlighted in orange. Values are highlighted red if below 1 (below 0 for log values) and highlighted blue if above 1 (above 0 for log values).

A.1 NMT tree

TABLE A.1: NMT Group 1.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’Oleu_XP_022886830.1’	14.3178	9.3856	1.1559	0.9725
’Solyc11g072870.1.1’	8.2069	5.5874	0.9142	0.7472
GSVIVT01008379001	3.0234	2.4839	0.4805	0.3951
GSVIVT01008381001	2.4886	2.0281	0.3960	0.3071
GSVIVT01008380001	0.9471	1.0763	-0.0236	0.0319
’Prupe.5G228500.1.p’	3.3857	2.5719	0.5296	0.4103
’Prupe.5G228600.1.p’	4.8937	2.9688	0.6896	0.4726
’Prupe.5G228700.1.p’	4.8179	2.7327	0.6829	0.4366
’Prupe.5G229000.1.p’	13.5158	6.2972	1.1308	0.7991
’Prupe.5G228900.1.p’	3.6311	2.5302	0.5600	0.4032
’Prupe.5G228800.1.p’	2.6064	2.0969	0.4160	0.3216
’Prupe.8G008300.1.p’	4.3437	3.1643	0.6379	0.5003
’Cac_g033654.t2’	0.2395	0.3282	-0.6207	-0.4839
’Cac_g031894.t1’	12.2915	7.0752	1.0896	0.8497
GWHPAAAL006898	3.6316	2.6532	0.5601	0.4238
’Potri.015G041900.1’	6.4108	3.8698	0.8069	0.5877
’Potri.012G052400.1’	0.0513	0.3218	-1.2895	-0.4924
’Potri.012G049900.1’	0.0513	0.3218	-1.2895	-0.4924
’Prupe.5G228400.1.p’	3.2310	2.3704	0.5093	0.3748
GWHPAAAL006899	2.6822	2.2538	0.4285	0.3529
Mean	4.7384	3.1059	0.3882	0.3459
Median	3.5084	2.5510	0.5448	0.4067

TABLE A.2: NMT Group 2.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’Ceug007g028170.1’	0.0991	0.1215	-1.0038	-0.9153
’chr_E_sg_C_Cara009g022920.1’	0.5584	0.4006	-0.2530	-0.3973
’chr_E_sg_E_Cara010g024030.1’	0.1031	0.3868	-0.9868	-0.4125
’Ccan005g025570.1’	0.1031	0.3868	-0.9868	-0.4125
Mean	0.2159	0.3240	-0.8076	-0.5344
Median	0.1031	0.3868	-0.9868	-0.4125

TABLE A.3: NMT Group 3.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’AT5G56300.1’	4.1991	5.0076	0.6232	0.6996
’AT4G26420.1’	5.0739	5.8063	0.7053	0.7639
’AT4G26460.1’	33.6399	28.4902	1.5269	1.4547
GSVIVT01025514001	2.2083	3.4136	0.3441	0.5332
’Nenu_XP_010276463.1’	3.7792	4.4307	0.5774	0.6465
’evm_27.model.AmTr_v1.0_scaffold00119.22’	10.6134	7.9131	1.0259	0.8983
Mean	9.9190	9.1769	0.8004	0.8327
Median	4.6365	5.4069	0.6643	0.7318

TABLE A.4: NMT Group 4.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
’Gese_maker-scaffold_250-augustus-gene-1.57-mRNA-1’	23.6165	19.1821	1.3732	1.2829
’Cac_g006957.t1’	19.7120	15.7930	1.2947	1.1985
’Ceug005g007070.1’	29.6872	23.1950	1.4726	1.3654
’Prupe.1G159300.1.p’	15.9823	15.9823	1.2036	1.2036
Mean	22.2495	18.5381	1.3360	1.2626
Median	21.6642	17.5822	1.3340	1.2433

TABLE A.5: NMT Group 5.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
’chr_J_sg_C_Cara019g007520.1’	0.7369	1.0284	-0.1326	0.0122
’Ceug083g007870.1’	0.4241	0.8434	-0.3725	-0.0740
’Ccan010g007360.1’	0.1323	0.3555	-0.8786	-0.4492
’chr_J_sg_E_Cara020g007440.1’	1.0366	0.9015	0.0156	-0.0450
Mean	0.5825	0.7822	-0.3420	-0.1390
Median	0.5805	0.8725	-0.2526	-0.0595

TABLE A.6: NMT Group 6.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
’Ccan003g012340.1’	0.0771	0.4805	-1.1132	-0.3183
’Ccan003g012350.1’	0.7105	0.9382	-0.1484	-0.0277
’Ccan003g012370.1’	0.7274	0.9467	-0.1382	-0.0238
’Ceug110g013400.1’	7.2348	4.8105	0.8594	0.6822
’chr_C_sg_E_Cara006g012490.1’	0.7130	0.9570	-0.1469	-0.0191
’chr_C_sg_E_Cara006g012460.1’	0.0855	0.4855	-1.0680	-0.3138
’Ceug110g013380.1’	0.0667	0.4022	-1.1760	-0.3956
Mean	1.3736	1.2887	-0.4188	-0.0594
Median	0.7105	0.9382	-0.1484	-0.0277

TABLE A.7: NMT Group 7.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
CRO_T127820	34.1836	32.6863	1.5338	1.5144
Thecc1EG006847t1	35.5406	33.1110	1.5507	1.5200
Thecc1EG006849t1	1.9811	5.1442	0.2969	0.7113
’Aqcoe6G125700.1.p’	4.4428	5.9675	0.6477	0.7758
’Ceug005g005880.1’	36.0962	36.0962	1.5575	1.5575
Mean	22.4489	22.6010	1.1173	1.2158
Median	34.1836	32.6863	1.5338	1.5144

TABLE A.8: NMT Group 8.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’Ceug027g013710.1’	7.8143	8.6759	0.8929	0.9383
’chr_I_sg_E_Cara018g012360.1’	0.0961	2.8278	-1.0174	0.4515
’Ceug083g016650.1’	4.5933	6.5419	0.6621	0.8157
’Ccan009g012130.1’	0.1046	2.9901	-0.9806	0.4757
’chr_I_sg_C_Cara017g011640.1’	0.6108	3.4031	-0.2141	0.5319
’Ceug083g016540.1’	0.4456	2.3274	-0.3511	0.3669
’chr_I_sg_E_Cara018g012430.1’	0.9305	2.9808	-0.0313	0.4743
’Ccan009g012120.1’	2.1801	4.4021	0.3385	0.6437
’Ceug027g013700.1’	0.4436	3.7616	-0.3530	0.5754
’Ceug083g016640.1’	24.0031	22.2251	1.3803	1.3468
’chr_I_sg_E_Cara018g012370.1’	3.2961	4.6606	0.5180	0.6684
’chr_I_sg_E_Cara018g012450.1’	2.4193	4.1380	0.3837	0.6168
’Ccan009g012030.1’	0.3959	3.7650	-0.4025	0.5758
’chr_I_sg_C_Cara017g011550.1’	0.3989	3.7674	-0.3992	0.5760
’Ceug083g016580.1’	0.1278	3.5257	-0.8935	0.5472
’chr_I_sg_E_Cara018g012400.1’	0.3929	3.7333	-0.4057	0.5721
’Ccan009g012070.1’	0.4431	4.0448	-0.3535	0.6069
’chr_I_sg_C_Cara017g011590.1’	0.1431	3.8152	-0.8444	0.5815
’Ccan009g012100.1’	0.4626	4.1080	-0.3348	0.6136
’Cara4866g005000.1’	0.1423	3.8639	-0.8469	0.5870
’chr_I_sg_C_Cara017g011630.1’	4.3104	6.1492	0.6345	0.7888
’Ceug083g016620.1’	0.2731	2.8643	-0.5637	0.4570
’Ceug035g010510.1’	0.5022	4.2645	-0.2991	0.6299
’chr_I_sg_E_Cara018g012380.1’	0.4665	4.2375	-0.3312	0.6271
’Cara389g005030.1’	2.6476	4.8518	0.4229	0.6859
’Cara389g005010.1’	0.0919	2.7250	-1.0368	0.4354
’Ceug005g006130.1’	4.5059	5.9232	0.6538	0.7726
’chr_A_sg_E_Cara002g005590.1’	0.0844	2.5018	-1.0737	0.3983
’Cara389g005040.1’	17.4135	17.1625	1.2409	1.2346
’Cara383g005040.1’	22.7638	21.7135	1.3572	1.3367
’chr_A_sg_C_Cara001g005940.1’	0.4081	2.7213	-0.3892	0.4348
’Ccan001g005960.1’	0.8744	2.8124	-0.0583	0.4491
’chr_I_sg_E_Cara018g012460.1’	0.2413	2.2912	-0.6174	0.3601
’Ceug083g016610.1’	24.0805	22.9862	1.3817	1.3615
’Ccan009g012020.1’	0.1188	3.3584	-0.9252	0.5261
’chr_I_sg_C_Cara017g011540.1’	0.9088	3.9850	-0.0416	0.6004
’Ceug083g016530.1’	2.2006	3.5288	0.3426	0.5476
’chr_B_sg_E_Cara004g013960.1’	0.1037	2.5284	-0.9844	0.4028
’Ceug044g015570.1’	0.4281	2.6716	-0.3685	0.4268
’Ccan002g014260.1’	0.9536	3.0238	-0.0207	0.4805
’chr_B_sg_C_Cara003g014350.1’	1.4592	3.2296	0.1641	0.5091
’Ceug044g015550.1’	1.4438	3.2047	0.1595	0.5058
’chr_B_sg_E_Cara004g013990.1’	0.1150	2.6555	-0.9394	0.4241
Mean	3.1590	5.3709	-0.1057	0.6270
Median	0.4665	3.7616	-0.3312	0.5754

TABLE A.9: NMT Group 9.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’TEA024636.1’	14.3281	13.1053	1.1562	1.1174
’Rico_XP_002534362.2’	9.8526	9.7350	0.9936	0.9883
’Potri.014G168200.1’	8.3159	8.7133	0.9199	0.9402
’Rico_XP_025015748.1’	7.8078	8.5555	0.8925	0.9322
’Rico_XP_002537751.2’	15.5102	13.6778	1.1906	1.1360
Mean	11.1629	10.7574	1.0306	1.0228
Median	9.8526	9.7350	0.9936	0.9883

TABLE A.10: NMT Group 10.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
Thecc1EG042578t1	0.3642	2.5677	-0.4387	0.4095
Thecc1EG042580t1	0.5518	3.1290	-0.2582	0.4954
Thecc1EG042584t1	2.5585	3.8193	0.4080	0.5820
Thecc1EG026299t1	5.5035	6.8428	0.7406	0.8352
Thecc1EG042577t1	41.8472	39.3704	1.6217	1.5952
Thecc1EG042587t1	0.0707	4.5426	-1.1504	0.6573
Thecc1EG042590t2	0.0707	4.5426	-1.1504	0.6573
Mean	7.2809	9.2592	-0.0325	0.7474
Median	0.5518	4.5426	-0.2582	0.6573

Table A.11: NMT Group 11.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'chr_G_sg_E_Cara014g014730.1'	0.7900	1.2704	-0.1024	0.1040
'Ceug046g015300.1'	0.1318	0.8929	-0.8802	-0.0492
'Ccan007g014410.1'	0.1266	0.1520	-0.8976	-0.8181
'chr_G_sg_C_Cara013g013100.1'	0.4272	0.3315	-0.3694	-0.4795
Mean	0.3689	0.6617	-0.5624	-0.3107
Median	0.2795	0.6122	-0.6248	-0.2644

Table A.12: NMT Group 12.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'Solyc02g084950.2.1'	8.9171	8.9171	0.9502	0.9502
'Oleu_XP_022883765.1'	5.5931	6.3919	0.7477	0.8056
'Sein_XP_011083194.1'	10.5364	10.8205	1.0227	1.0342
'Sein_XP_011091223.1'	16.3596	15.8214	1.2138	1.1992
'Utgi_unitig_748.g8672.t1'	19.0860	18.1412	1.2807	1.2587
Mean	12.0984	12.0184	1.0430	1.0496
Median	10.5364	10.8205	1.0227	1.0342

Table A.13: NMT Group 13.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'Sein_XP_020547784.1'	16.2398	12.2563	1.2106	1.0884
'Prupe.1G321300.1.p'	6.6235	6.3471	0.8211	0.8026
'cal_g010094.t1'	10.6111	8.4447	1.0258	0.9266
'cal_g010097.t1'	8.4782	7.2940	0.9283	0.8630
Achn373691	4.8857	5.2576	0.6889	0.7208
'Nenu_XP_010243761.1'	3.8923	3.8923	0.5902	0.5902
Mean	8.4551	7.2487	0.8775	0.8319
Median	7.5508	6.8206	0.8747	0.8328

Table A.14: NMT Group 14.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'Cac_g003539.t1'	4.2813	3.9442	0.6316	0.5960
'TEA007272.1'	8.5237	6.4344	0.9306	0.8085
Thecc1EG019319t1	6.7596	5.7522	0.8299	0.7598
'Potri.008G127000.1'	5.9120	5.0553	0.7717	0.7038
'Cac_g019406.t1'	6.1532	5.2121	0.7891	0.7170
'Nenu_XP_010243764.1'	2.8956	2.8956	0.4617	0.4617
Mean	5.7542	4.8823	0.7358	0.6745
Median	6.0326	5.1337	0.7804	0.7104

TABLE A.15: NMT Group 15.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
’Ccan011g018210.1’	1.4448	5.3435	0.1598	0.7278
’Cara750g005010.1’	4.5333	7.2830	0.6564	0.8623
’chr_K_sg_C_Cara021g015470.1’	1.0865	5.9786	0.0360	0.7766
’Ceug035g022220.1’	0.8426	5.5518	-0.0744	0.7444
’chr_K_sg_E_Cara022g015960.1’	0.8894	5.5715	-0.0509	0.7460
’chr_G_sg_C_Cara013g021370.1’	15.0424	15.2744	1.1773	1.1840
’chr_G_sg_E_Cara014g023280.1’	7.4892	8.8655	0.8744	0.9477
’chr_G_sg_C_Cara013g021380.1’	0.0678	3.0935	-1.1687	0.4904
’Ccan007g023250.1’	0.0827	3.6656	-1.0826	0.5641
’Ceug005g025460.1’	0.1704	6.1487	-0.7685	0.7888
’Ceug046g026100.1’	3.1269	8.0781	0.4951	0.9073
’Ceug046g026820.1’	1.5557	5.5929	0.1919	0.7476
’Ceug005g025480.1’	0.7630	5.6305	-0.1174	0.7505
’Ccan007g023280.1’	0.1311	5.1837	-0.8824	0.7146
’chr_G_sg_C_Cara013g021400.1’	0.6663	3.9843	-0.1763	0.6004
’Ceug005g025470.1’	0.6438	3.8686	-0.1913	0.5875
’Ccan007g023260.1’	0.1072	4.4385	-0.9700	0.6472
’chr_G_sg_C_Cara013g021390.1’	0.5148	6.2701	-0.2883	0.7973
’Cara4454g005000.1’	1.4665	6.8970	0.1663	0.8387
’Ccan007g023240.1’	0.6878	4.4190	-0.1625	0.6453
’Ceug046g026790.1’	0.6363	6.1758	-0.1963	0.7907
’chr_G_sg_E_Cara014g023260.1’	0.6519	6.1862	-0.1858	0.7914
’Ceug046g026800.1’	1.1760	6.1361	0.0704	0.7879
’Ceug046g026850.1’	0.4662	5.6436	-0.3314	0.7516
’Ceug005g025490.1’	0.3879	4.7860	-0.4113	0.6800
’chr_G_sg_C_Cara013g021420.1’	1.1426	6.9175	0.0579	0.8400
’Ccan007g023310.1’	0.1788	6.2965	-0.7477	0.7991
’chr_G_sg_E_Cara014g023300.1’	0.7111	2.9675	-0.1481	0.4724
’Ceug046g026110.1’	1.1281	3.1737	0.0523	0.5016
Mean	1.6480	5.8421	-0.1385	0.7408
Median	0.7111	5.6305	-0.1481	0.7505

TABLE A.16: NMT Group 16.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
’chr_G_sg_E_Cara014g013270.1’	0.6245	4.7979	-0.2045	0.6810
’Ceug046g017120.1’	0.1516	5.7849	-0.8193	0.7623
’Ceug046g013720.1’	1.3009	6.2857	0.1142	0.7984
’chr_G_sg_C_Cara013g011520.1’	1.3692	5.6419	0.1365	0.7514
’Ccan007g012920.1’	0.5904	5.2428	-0.2288	0.7196
’Ceug007g006290.1’	0.4441	9.9228	-0.3525	0.9966
’chr_E_sg_E_Cara010g007130.1’	0.4232	9.9092	-0.3734	0.9960
’Ccan005g008070.1’	0.1399	10.0902	-0.8540	1.0039
’chr_E_sg_C_Cara009g007210.1’	0.4469	10.2860	-0.3498	1.0122
’Ccan003g016320.1’	0.1425	10.1339	-0.8461	1.0058
’Ceug110g020240.1’	0.4383	10.3220	-0.3583	1.0138
’chr_C_sg_E_Cara006g016090.1’	0.4658	7.4890	-0.3318	0.8744
’Cara1418g005000.1’	0.0734	6.3121	-1.1341	0.8002
’chr_C_sg_C_Cara005g016040.1’	30.5184	29.8572	1.4846	1.4750
Mean	2.6521	9.4340	-0.2941	0.9208
Median	0.4455	8.6991	-0.3512	0.9352

TABLE A.17: NMT Group 17.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
’Ceug005g015230.1’	3.6448	4.2740	0.5617	0.6308
’Ceug005g015260.1’	0.1021	2.3912	-0.9908	0.3786
’Cara4701g005000.1’	0.7814	2.8025	-0.1071	0.4475
’chr_A_sg_E_Cara002g009760.1’	0.0788	2.0208	-1.1032	0.3055
’chr_A_sg_E_Cara002g009810.1’	1.4266	2.9193	0.1543	0.4653
’Ccan001g013300.1’	1.3566	2.6853	0.1325	0.4290
’chr_A_sg_C_Cara001g009810.1’	0.6585	2.1911	-0.1815	0.3407
’chr_A_sg_C_Cara001g009900.1’	0.7317	1.4088	-0.1357	0.1489
’chr_A_sg_E_Cara002g009890.1’	0.0463	0.1610	-1.3345	-0.7930
’chr_A_sg_E_Cara002g009900.1’	0.3249	0.3117	-0.4882	-0.5063
’Ceug035g011850.1’	0.0976	0.2273	-1.0104	-0.6433
’Ceug007g023110.1’	0.3020	0.3027	-0.5200	-0.5190
Mean	0.7960	1.8080	-0.4186	0.0570
Median	0.4917	2.1059	-0.3348	0.3231

TABLE A.18: NMT Group 18.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’chr_G_sg_C_Cara013g021880.1’	0.0450	0.0521	-1.3463	-1.2833
’Ccan007g023880.1’	0.3763	0.2436	-0.4244	-0.6134
’Ceug046g027740.1’	0.0508	1.6141	-1.2938	0.2079
’chr_G_sg_E_Cara014g024140.1’	0.0517	1.6145	-1.2863	0.2080
Mean	0.1310	0.8811	-1.0877	-0.3702
Median	0.0513	0.9288	-1.2901	-0.2027

TABLE A.19: NMT Group 19.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’AT1G68040.1’	6.6168	4.2661	0.8206	0.6300
’Prupe.1G334700.1.p’	3.3476	2.5448	0.5247	0.4057
Thecc1EG011286t1	3.2474	2.4813	0.5115	0.3947
Thecc1EG011287t1	5.8986	3.3284	0.7708	0.5222
’Potri.008G136200.1’	7.0403	3.1883	0.8476	0.5036
’Rico_XP_002526004.1’	7.5088	3.3219	0.8756	0.5214
’Potri.008G136300.1’	0.5113	0.9618	-0.2914	-0.0169
’Potri.010G104700.1’	5.3463	3.8874	0.7281	0.5897
Thecc1EG011290t1	0.9862	0.9862	-0.0060	-0.0060
Mean	4.5004	2.7740	0.5313	0.3938
Median	5.3463	3.1883	0.7281	0.5036

TABLE A.20: NMT Group 20.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’chr_K_sg_C_Cara021g013540.1’	0.2327	0.1948	-0.6332	-0.7103
’Ccan011g016030.1’	0.1334	0.1413	-0.8749	-0.8500
’Ceug035g019770.1’	0.0742	0.1275	-1.1293	-0.8943
’chr_K_sg_E_Cara022g014200.1’	0.0213	0.0991	-1.6717	-1.0038
Mean	0.1154	0.1407	-1.0773	-0.8646
Median	0.1038	0.1344	-1.0021	-0.8722

TABLE A.21: NMT Group 21.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’Prupe.1G345700.1.p’	0.5201	0.5201	-0.2839	-0.2839
CRO_T136128	1.9522	1.4241	0.2905	0.1535
’chr_K_sg_C_Cara021g013540.1’	0.2327	0.7085	-0.6332	-0.1497
’Ccan011g016030.1’	0.1334	0.6834	-0.8749	-0.1653
’Ceug035g019770.1’	0.0742	0.6692	-1.1293	-0.1744
’chr_K_sg_E_Cara022g014200.1’	0.0213	0.6559	-1.6717	-0.1832
’TEA004178.1’	1.6772	1.2751	0.2246	0.1055
’TEA013347.1’	1.5649	1.2187	0.1945	0.0859
’Cac_g020334.t1’	0.5498	0.6303	-0.2598	-0.2005
Mean	0.7473	0.8650	-0.4604	-0.0902
Median	0.5201	0.6834	-0.2839	-0.1653

A.2 FAD tree

TABLE A.22: FAD Group 1.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
Ceug110g018420.1	1.1983	9.9526	0.0786	0.9979
chr_A_sg_E_Cara002g008860.1	1.1983	9.9526	0.0786	0.9979
chr_A_sg_C_Cara001g009000.1	1.1838	9.9437	0.0733	0.9975
Ccan001g012300.1	3.6632	11.1751	0.5639	1.0483
Ccan2215g005060.1	13.2528	73.7347	1.1223	1.8677
chr_A_sg_C_Cara001g008950.1	76.3799	113.5802	1.8830	2.0553
chr_A_sg_C_Cara001g008980.1	9.3523	106.4704	0.9709	2.0272
Ccan001g012340.1	0.5278	104.5390	-0.2775	2.0193
chr_A_sg_E_Cara002g008850.1	2.7153	69.1950	0.4338	1.8401
Ceug110g018410.1	2.6530	69.1643	0.4237	1.8399
Ccan001g012190.1	130.1246	131.5642	2.1144	2.1191
chr_A_sg_E_Cara002g008820.1	72.9730	107.6926	1.8632	2.0322
Ceug110g018360.1	3.1955	77.8203	0.5045	1.8911
Ceug005g014180.1	27.2322	82.1720	1.4351	1.9147
chr_A_sg_E_Cara002g008810.1	68.9331	102.3509	1.8384	2.0101
chr_A_sg_C_Cara001g008840.1	0.1662	67.9211	-0.7793	1.8320
Ccan001g012170.1	2.4715	69.0565	0.3930	1.8392
Ceug005g014150.1	0.1504	67.4861	-0.8226	1.8292
Cara3609g005000.1	0.0135	88.9431	-1.8690	1.9491
chr_A_sg_E_Cara002g008760.1	0.0135	88.9431	-1.8690	1.9491
Ccan001g012150.1	0.0070	67.8877	-2.1538	1.8318
Ccan001g012090.1	0.007017802241	67.88774086	-2.153798874	1.831791357
Ceug005g014130.1	31.04209964	61.08327651	1.491951089	1.785922325
Ccan001g012130.1	18.07050121	69.67364507	1.256970198	1.843068532
chr_A_sg_C_Cara001g008820.1	203.0053615	172.063684	2.307507508	2.235689217
Mean	26.7812	76.0101	0.3563	1.7834
Median	2.7153	69.6736	0.4338	1.8431

TABLE A.23: FAD Group 2.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
evm_27.model.AmTr_v1.0_scaffold00025.315	100.7689	68.8855	2.0033	1.8381
Solyc07g056110.2.1	228.8932	107.4802	2.3596	2.0313
AT3G13320.1	134.6376	56.7864	2.1292	1.7542
AT1G55720.1	316.3449	76.2598	2.5002	1.8823
AT1G55730.1	94.4115	42.9800	1.9750	1.6333
Oleu_XP_022843877.1	65.8711	45.4613	1.8187	1.6576
Utgi_unitig_748.g7078.t1	212.7246	98.5450	2.3278	1.9936
Achn164281	138.1034	70.8867	2.1402	1.8506
Achn301291	62.3290	43.9942	1.7947	1.6434
Achn226261	117.8471	85.8365	2.0713	1.9337
Thecc1EG008118t1	92.3272	58.5755	1.9653	1.7677
evm_27.model.AmTr_v1.0_scaffold00049.235	486.6901	179.4147	2.6873	2.2539
Potri.013G130300.1	215.0006	92.1039	2.3324	1.9643
Nenu_XP_010249695.1	1044.2278	461.9859	3.0188	2.6646
Achn157911	1617.9146	705.0604	3.2090	2.8482
AT2G36100.1	448.1088	142.6823	2.6514	2.1544
Thecc1EG041359t1	505.3685	133.1967	2.7036	2.1245
Achn200321	423.7515	116.0082	2.6271	2.0645
chr_K_sg_E_Cara022g014450.1	766.7785	451.7005	2.8847	2.6549
chr_E_sg_C_Cara009g019320.1	856.9218	505.4579	2.9329	2.7037
Gese_maker-scaffold_347-snap-gene-0.69-mRNA-1	300.6091	100.3610	2.4780	2.0016
Achn366671	583.0734	170.7424	2.7657	2.2323
Potri.006G059800.1	910.4581	475.7196	2.9593	2.6774
GWHPAAAL007328	492.8409	281.5811	2.6927	2.4496
Oleu_XP_022847770.1	1115.5011	494.7743	3.0475	2.6944
Achn095091	1156.4147	511.9351	3.0631	2.7092
Ccan008g024740.1	60.4051	52.2227	1.7811	1.7179
CRO_T131995	69.7530	52.6778	1.8436	1.7216
cal_g006956.t1	91.0120	58.4475	1.9591	1.7668
Cac_g011350.t1	40.3824	42.8948	1.6062	1.6324
Mean	424.9824	192.8219	2.4110	2.1007
Median	264.7511	99.4530	2.4188	1.9976

TABLE A.24: FAD Group 3.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
Cara630g005010.1	0.1486	0.1627	-0.8279	-0.7887
chr_B_sg_E_Cara004g018300.1	8.3751	2.4067	0.9230	0.3814
Ceug044g019850.1	20.7571	5.7738	1.3172	0.7615
Ccan002g018220.1	0.1927	0.1837	-0.7152	-0.7359
chr_B_sg_C_Cara003g018210.1	175.3080	175.3080	2.2438	2.2438
Mean	40.9563	36.7670	0.5882	0.3724
Median	8.3751	2.4067	0.9230	0.3814

TABLE A.25: FAD Group 4.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
chr_B_sg_E_Cara004g011390.1	0.4717	12.0063	-0.3264	1.0794
Ceug044g011900.1	0.0377	15.9661	-1.4239	1.2032
Ceug044g011800.1	0.0377	15.9661	-1.4239	1.2032
chr_B_sg_C_Cara003g011280.1	20.3014	18.2977	1.3075	1.2624
Ccan002g011230.1	13.3101	14.3941	1.1242	1.1582
Mean	6.8317	15.3261	-0.1485	1.1813
Median	0.4717	15.9661	-0.3264	1.2032

TABLE A.26: FAD Group 5.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
Prupe.4G054000.1.p	1514.5822	1511.8029	3.1803	3.1795
Prupe.1G346400.1.p	628.2662	1024.3499	2.7981	3.0104
Potri.005G093600.1	902.2470	1173.8644	2.9553	3.0696
Utgi_unitig_52.g17488.t1	903.8435	1200.9965	2.9561	3.0795
Sein_XP_011071842.2	631.9249	1064.5825	2.8007	3.0272
Ceug097g021090.1	751.6655	1074.9981	2.8760	3.0314
evm_27.model.AmTr_v1.0_scaffold00025.9	1484.1549	1484.1549	3.1715	3.1715
Mean	973.8120	1219.2499	2.9626	3.0813
Median	902.2470	1173.8644	2.9553	3.0696

TABLE A.27: FAD Group 6.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
AT1G21722.1	1690.5646	1690.5646	3.2280	3.2280
Nenu_XP_010245027.1	1614.0149	1671.3433	3.2079	3.2231
Solyc07g006400.1.1	877.0058	1356.2560	2.9430	3.1323
CRO_T115871	1460.9730	1657.7052	3.1646	3.2195
Mean	1410.6396	1593.9673	3.1359	3.2007
Median	1537.4940	1664.5242	3.1863	3.2213

TABLE A.28: FAD Group 7.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
Ceug021g007580.1	0.04965	0.63023	-1.30405	-0.20050
Ccan008g008060.1	0.04965	0.63023	-1.30405	-0.20050
chr_H_sg_C_Cara015g007010.1	0.04966	0.68498	-1.30402	-0.16432
chr_H_sg_E_Cara016g007070.1	0.04966	0.68498	-1.30402	-0.16432
Mean	0.0497	0.6576	-1.3040	-0.1824
Median	0.0497	0.6576	-1.3040	-0.1824

TABLE A.29: FAD Group 8.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
GSVIVT01008198001	1131.0462	1020.4353	3.0535	3.0088
Aqcoe3G387600.1.p	668.4800	667.4789	2.8251	2.8244
Rico_XP_015584138.1	1050.7591	1035.4166	3.0215	3.0151
Solyc12g038410.1.1	1627.1741	1538.9464	3.2114	3.1872
Mean	1119.3648	1065.5693	3.0279	3.0089
Median	1090.9027	1027.9259	3.0375	3.0120

TABLE A.30: FAD Group 9.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
AT4G22850.1	1809.4979	1496.6685	3.2576	3.1751
GSVIVT01028650001	1360.9930	1160.7931	3.1339	3.0648
Ipni_XP_019154785.1	1496.7819	1224.2715	3.1752	3.0879
GWHPAAAL011115	1635.0251	1289.7102	3.2135	3.1105
Oleu_XP_022845037.1	2220.5526	1686.4948	3.3465	3.2270
AT2G18630.1	1439.4406	992.3566	3.1582	2.9967
Utgi_unitig_22.g5413.t1	1337.9605	942.6885	3.1264	2.9744
Gese_maker-scaffold_0-snap-gene-46.57-mRNA-1	1622.6824	1332.6702	3.2102	3.1247
Thecc1EG042507t1	1204.0800	1104.1143	3.0807	3.0430
chr_B_sg_E_Cara004g007540.1	1.4557	562.3953	0.1631	2.7500
Ceug044g007790.1	36.7045	571.5439	1.5647	2.7570
chr_B_sg_C_Cara003g007660.1	98.5249	519.4493	1.9935	2.7155
chr_J_sg_E_Cara020g011150.1	991.1451	871.0033	2.9961	2.9400
Aqcoe5G256600.1.p	1234.6662	988.4856	3.0915	2.9950
chr_F_sg_E_Cara012g009600.1	1623.8453	1436.5178	3.2105	3.1573
evm_27.model.AmTr_v1.0_scaffold00004.116	1992.7639	1704.8687	3.2995	3.2317
Prupe.7G240600.1.p	1475.7234	1308.4798	3.1690	3.1168
Mean	1269.5202	1128.9713	2.8347	3.0275
Median	1439.4406	1160.7931	3.1582	3.0648

TABLE A.31: FAD Group 10.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
Solyc12g008950.1.1	458.9090	849.3119	2.6617	2.9291
Rico_XP_002521382.2	636.4089	990.5391	2.8037	2.9959
chr_F_sg_E_Cara012g016050.1	1030.2218	1227.6600	3.0129	3.0891
Sein_XP_011093753.1	707.0679	1103.1315	2.8495	3.0426
Cac_g001425.t1	532.0437	1022.4027	2.7259	3.0096
Cac_g006743.t1	1421.8578	1426.5133	3.1529	3.1543
Prupe.4G101100.1.p	995.9929	995.9929	2.9983	2.9983
Mean	826.0717	1087.9359	2.8864	3.0313
Median	707.0679	1022.4027	2.8495	3.0096

TABLE A.32: FAD Group 11.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
GWHPAAAL011392	515.4394	1250.9065	2.7122	3.0972
Ceug044g016850.1	20.0565	1265.7804	1.3023	3.1024
chr_B_sg_C_Cara003g015530.1	142.7868	1369.7827	2.1547	3.1367
Ccan002g015410.1	16.0172	1350.4019	1.2046	3.1305
Cara8374g005030.1	1684.0502	1531.6034	3.2264	3.1851
chr_J_sg_E_Cara020g012700.1	1841.7799	1618.1368	3.2652	3.2090
Potri.T148800.1	1696.1758	1499.5017	3.2295	3.1759
Mean	845.1865	1412.3019	2.4421	3.1481
Median	515.4394	1369.7827	2.7122	3.1367

TABLE A.33: FAD Group 12.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
Ccan002g050620.1	29.4723	68.7715	1.4694	1.8374
chr_B_sg_E_Cara004g020970.1	2059.0117	1781.0197	3.3137	3.2507
chr_B_sg_C_Cara003g048060.1	0.6661	39.8066	-0.1765	1.6000
Ceug110g021890.1	37.4168	33.4348	1.5731	1.5242
chr_B_sg_E_Cara004g048870.1	1.0484	13.5556	0.0205	1.1321
Mean	425.5231	387.3176	1.2400	1.8689
Median	29.4723	39.8066	1.4694	1.6000

TABLE A.34: FAD Group 13.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
Ceug021g025500.1	71.6542	88.9406	1.8552	1.9491
chr_H_sg_E_Cara016g019020.1	33.1726	66.8189	1.5208	1.8249
chr_H_sg_C_Cara015g018680.1	46.8118	43.3978	1.6704	1.6375
Ccan008g022210.1	30.7269	33.9689	1.4875	1.5311
Mean	45.5913	58.2816	1.6335	1.7356
Median	39.9922	55.1084	1.5956	1.7312

TABLE A.35: FAD Group 14.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
chr_B_sg_C_Cara003g015340.1	733.0588	928.7202	2.8651	2.9679
Ipni_XP_019151293.1	468.0221	772.4780	2.6703	2.8879
Nenu_XP_010268179.1	919.9183	1085.0169	2.9637	3.0354
GWHPAAAL024740	23.5172	739.2566	1.3714	2.8688
GWHPAAAL024741	792.7379	1095.9520	2.8991	3.0398
Nenu_XP_010247465.1	1397.1343	1392.2182	3.1452	3.1437
Ipni_XP_019185825.1	1082.1529	1172.4421	3.0343	3.0691
AT5G66675.1	1313.5362	1328.1255	3.1184	3.1232
Oleu_XP_022867205.1	229.9841	663.3518	2.3617	2.8217
GWHPAAAL014977	930.2214	1123.4014	2.9686	3.0505
Prupe.4G276200.1.p	889.3531	1100.9832	2.9491	3.0418
chr_D_sg_C_Cara007g018630.1	344.8201	701.3620	2.5376	2.8459
Oleu_XP_022899642.1	1496.5500	1478.3133	3.1751	3.1698
GWHPAAAL016173	1067.2422	793.2735	3.0283	2.8994
Sein_XP_011086820.1	251.4932	296.9622	2.4005	2.4727
Prupe.1G143600.1.p	1629.5439	1422.5366	3.2121	3.1531
chr_F_sg_E_Cara012g024260.1	1635.1957	1394.2113	3.2136	3.1443
Oleu_XP_022846830.1	1665.3790	1418.6563	3.2215	3.1519
Nenu_XP_010270454.1	1559.5420	1334.4334	3.1930	3.1253
Prupe.3G260800.1.p	2041.5447	1725.4051	3.3100	3.2369
Rico_XP_002520992.1	447.0637	422.8315	2.6504	2.6262
Achn018541	547.6470	486.5622	2.7385	2.6871
Gese_maker-scaffold_35-augustus-gene-5.40-mRNA-1	849.7991	664.4624	2.9293	2.8225
Utgi_unitig_747.g21347.t1	1479.7858	1052.3054	3.1702	3.0221
Ccan004g005290.1	1716.1545	1503.5173	3.2346	3.1771
Mean	1020.4559	1043.8711	2.8945	2.9834
Median	930.2214	1095.9520	2.9686	3.0398

A.3 Terpene tree

TABLE A.36: Terpene Group 1.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
'Ccan002g047400.1'	0.0579	0.1564	-1.2374	-0.8057
'chr_B_sg_C_Cara003g045290.1'	0.0563	0.1556	-1.2493	-0.8079
'Ceug044g052510.1'	0.2652	0.2164	-0.5765	-0.6647
'chr_B_sg_E_Cara004g045890.1'	0.0764	0.1221	-1.1167	-0.9134
'Ceug044g053090.1'	0.0606	0.0963	-1.2177	-1.0165
Mean	0.1033	0.1494	-1.0795	-0.8417
Median	0.0606	0.1556	-1.2177	-0.8079

TABLE A.37: Terpene Group 2.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
'chr_J_sg_C_Cara019g019360.1'	0.5007	1.7906	-0.3004	0.2530
'Ccan010g020940.1'	0.0481	1.5642	-1.3182	0.1943
'Ceug083g022060.1'	0.1502	1.6153	-0.8233	0.2083
'chr_J_sg_E_Cara020g020090.1'	0.3249	1.7026	-0.4883	0.2311
'chr_J_sg_C_Cara019g019300.1'	0.3628	2.5217	-0.4403	0.4017
'Ccan010g020890.1'	0.0270	2.3538	-1.5692	0.3718
'chr_J_sg_C_Cara019g019310.1'	0.0033	1.5624	-2.4847	0.1938
'Ccan010g020900.1'	0.0193	1.1846	-1.7154	0.0736
'Ceug083g022050.1'	0.0298	1.5801	-1.5265	0.1987
'chr_J_sg_E_Cara020g019920.1'	0.8622	2.1350	-0.0644	0.3294
Mean	0.2328	1.8010	-1.0731	0.2456
Median	0.0991	1.6590	-1.0707	0.2197

TABLE A.38: Terpene Group 3.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
'Solyc00g154480.1.1'	34.9915	18.9529	1.5440	1.2777
'Solyc02g079910.1.1'	40.5072	15.4453	1.6075	1.1888
'Ipni_XP_019171893.1'	22.9520	9.5936	1.3608	0.9820
'Sein_XP_011096130.1'	16.7435	10.8294	1.2238	1.0346
'Oleu_XP_022845587.1'	4.9562	4.9289	0.6951	0.6927
'Oleu_XP_022845588.1'	9.1754	6.3353	0.9626	0.8018
Mean	21.5543	11.0142	1.2323	0.9963
Median	19.8477	10.2115	1.2923	1.0083

TABLE A.39: Terpene Group 4.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'TEA009166.1'	0.0203	0.0192	-1.6927	-1.7156
'TEA026270.1'	1.6019	0.4141	0.2046	-0.3829
'TEA031095.1'	0.0546	0.0273	-1.2626	-1.5636
'TEA012490.1'	0.0265	0.0215	-1.5762	-1.6668
'TEA009866.1'	4.0668	4.0668	0.6093	0.6093
Mean	1.1540	0.9098	-0.7435	-0.9439
Median	0.0546	0.0273	-1.2626	-1.5636

TABLE A.40: Terpene Group 5.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
Thecc1EG046206t1	0.28663	5.93400	-0.54267	0.77335
Thecc1EG046209t1	0.01468	7.72580	-1.83341	0.88794
Thecc1EG007798t1	0.01468	7.72580	-1.83341	0.88794
'Prupe.3G222300.1.p'	2.10345	9.24974	0.32293	0.96613
'Prupe.3G222200.1.p'	0.00842	8.20222	-2.07489	0.91393
Mean	0.4856	7.7675	-1.1923	0.8859
Median	0.0147	7.7258	-1.8334	0.8879

TABLE A.41: Terpene Group 6.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
Thecc1EG031741t1	21.2795	13.1641	1.3280	1.1194
Thecc1EG032886t1	21.5754	10.5576	1.3340	1.0236
Thecc1EG032806t1	19.8919	8.0173	1.2987	0.9040
Thecc1EG032777t1	7.9797	5.6349	0.9020	0.7509
Thecc1EG043473t1	6.3083	6.3083	0.7999	0.7999
Mean	15.4070	8.7364	1.1325	0.9196
Median	19.8919	8.0173	1.2987	0.9040

TABLE A.42: Terpene Group 7.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'Ceug034g040520.1'	36.2104	1145.3529	1.5588	3.0589
'chr_F_sg_E_Cara012g032470.1'	39.4033	1147.0726	1.5955	3.0596
'Ccan006g034970.1'	37.3471	1309.2987	1.5723	3.1170
'chr_F_sg_C_Cara011g031400.1'	302.8758	1452.0347	2.4813	3.1620
Mean	103.9591	1263.4397	1.8020	3.0994
Median	38.3752	1228.1857	1.5839	3.0883

TABLE A.43: Terpene Group 8.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'Ceug110g031060.1'	22542.7224	20127.2246	4.3530	4.3038
'Ceug097g018330.1'	176.4587	338.4685	2.2466	2.5295
'Ccan003g024120.1'	459.5332	586.3482	2.6623	2.7682
'Ccan003g024190.1'	58.2441	328.6842	1.7653	2.5168
'Ceug110g031110.1'	19593.4654	16397.4159	4.2921	4.2148
'chr_C_sg_C_Cara005g021420.1'	28.6092	241.6542	1.4565	2.3832
'Ceug097g018300.1'	311.7477	510.8589	2.4938	2.7083
'chr_C_sg_E_Cara006g022060.1'	9085.3037	7909.4848	3.9583	3.8981
'Ceug035g008560.1'	15.6585	15.6585	1.1947	1.1947
Mean	5807.9714	5161.7553	2.7136	2.9464
Median	311.7477	510.8589	2.4938	2.7083

TABLE A.44: Terpene Group 9.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
’chr_F_sg_C_Cara011g031280.1’	690.7884	997.1778	2.8393	2.9988
’Ccan006g034850.1’	345.4078	790.2921	2.5383	2.8978
’Cara951g005000.1’	1611.9441	4729.1621	3.2073	3.6748
’chr_F_sg_E_Cara012g032360.1’	54.8575	3898.6951	1.7392	3.5909
Mean	675.7494	2603.8318	2.5811	3.2906
Median	518.0981	2447.9365	2.6888	3.2948

TABLE A.45: Terpene Group 10.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
’Ceug034g040310.1’	1294.3218	1203.1045	3.1120	3.0803
’chr_F_sg_E_Cara012g032340.1’	1307.4075	1211.2483	3.1164	3.0832
’Ccan006g034830.1’	413.1201	1402.9994	2.6161	3.1471
’chr_F_sg_C_Cara011g031270.1’	1508.4974	2056.7300	3.1785	3.3132
Mean	1130.8367	1468.5205	3.0058	3.1559
Median	1300.8647	1307.1239	3.1142	3.1151

TABLE A.46: Terpene Group 11.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
’Oleu_XP_022895811.1’	36.9004	5626.7590	1.5670	3.7503
’Oleu_XP_022887509.1’	41.2388	5628.4092	1.6153	3.7504
’Sein_XP_011093795.1’	25.5212	4783.0986	1.4069	3.6797
’Sein_XP_020547258.1’	43.8498	6343.4483	1.6420	3.8023
’Sein_XP_011085901.1’	1504.1220	6963.4000	3.1773	3.8428
Mean	330.3264	5869.0230	1.8817	3.7651
Median	41.2388	5628.4092	1.6153	3.7504

TABLE A.47: Terpene Group 12.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
’Ceug044g048090.1’	959.6589	13943.5284	2.9821	4.1444
’Ceug034g032030.1’	43.1863	13510.2506	1.6353	4.1307
’chr_B_sg_E_Cara004g042350.1’	1108.5490	11133.7411	3.0448	4.0466
’chr_B_sg_C_Cara003g041350.1’	1751.9020	12614.3891	3.2435	4.1009
’Ccan002g043550.1’	450.8589	11906.2997	2.6540	4.0758
’Ccan010g018720.1’	7349.1220	15103.1379	3.8662	4.1791
’chr_J_sg_C_Cara019g017460.1’	13460.6056	18616.1320	4.1291	4.2699
’Ceug005g030330.1’	267.8840	2741.6889	2.4279	3.4380
’chr_A_sg_E_Cara002g021450.1’	268.8581	2742.1240	2.4295	3.4381
’Ccan001g024890.1’	34.7494	2626.2883	1.5409	3.4193
’chr_A_sg_C_Cara001g019700.1’	266.2690	2730.2647	2.4253	3.4362
Mean	2360.1494	9787.9859	2.7617	3.8799
Median	450.8589	11906.2997	2.6540	4.0758

TABLE A.48: Terpene Group 13.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
CRO_T114690	10470.9876	11414.1980	4.0200	4.0574
CRO_T110182	26723.9272	24040.5433	4.4269	4.3809
CRO_T110174	13007.3060	11896.7964	4.1142	4.0754
CRO_T110169	16290.9116	14348.8203	4.2119	4.1568
Mean	16623.2831	15425.0895	4.1933	4.1677
Median	14649.1088	13122.8084	4.1631	4.1161

TABLE A.49: Terpene Group 14.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'Ceug007g009590.1'	1.6662	20.8730	0.2217	1.3196
'Ceug083g016130.1'	26.7493	35.6582	1.4273	1.5522
'chr_J_sg_E_Cara020g015020.1'	3.9069	28.0104	0.5918	1.4473
'Cara4041g005000.1'	0.8355	13.9348	-0.0780	1.1441
'chr_J_sg_C_Cara019g015190.1'	27.8141	24.1848	1.4443	1.3835
'Ccan010g015350.1'	130.4618	75.5493	2.1155	1.8782
Mean	31.9056	33.0351	0.9538	1.4542
Median	15.3281	26.0976	1.0096	1.4154

TABLE A.50: Terpene Group 15.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'evm_27.model.AmTr_v1.0_scaffold00003.315'	77.9467	47.7420	1.8918	1.6789
'evm_27.model.AmTr_v1.0_scaffold00066.195'	73.2158	45.3766	1.8646	1.6568
'evm_27.model.AmTr_v1.0_scaffold00003.312'	191.5788	172.2818	2.2823	2.2362
'evm_27.model.AmTr_v1.0_scaffold00035.30'	13.9747	83.4800	1.1453	1.9216
'evm_27.model.AmTr_v1.0_scaffold00003.313'	0.0688	76.5274	-1.1623	1.8838
'evm_27.model.AmTr_v1.0_scaffold00003.314'	144.2728	148.6289	2.1592	2.1721
Mean	83.5096	95.6728	1.3635	1.9249
Median	75.5812	80.0037	1.8782	1.9027

TABLE A.51: Terpene Group 16.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'Cac_g025125.t1'	350.1262	215.0487	2.5442	2.3325
'Cac_g025124.t1'	61.0569	70.4870	1.7857	1.8481
'Cac_g025123.t1'	227.9721	178.6345	2.3579	2.2520
'Cac_g026292.t1'	235.8159	196.8498	2.3726	2.2941
'TEA019347.1'	347.3477	210.2145	2.5408	2.3227
Achn226941	656.9993	267.9784	2.8176	2.4281
Achn111061	909.4665	352.2841	2.9588	2.5469
'TEA005451.1'	148.2205	103.0710	2.1709	2.0131
'TEA005162.1'	56.4691	66.3532	1.7518	1.8219
'TEA005163.1'	538.6578	339.1307	2.7313	2.5304
Mean	353.2132	200.0052	2.4032	2.2390
Median	291.5818	203.5322	2.4567	2.3084

TABLE A.52: Terpene Group 17.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'chr_B_sg_E_Cara004g017980.1'	34.9469	798.9441	1.5434	2.9025
'Ceug044g019530.1'	35.2610	799.1001	1.5473	2.9026
'chr_B_sg_C_Cara003g017850.1'	8.3273	785.8224	0.9205	2.8953
'Ccan002g017870.1'	18.3032	790.7772	1.2625	2.8981
'chr_B_sg_E_Cara004g018000.1'	290.3180	1461.5740	2.4629	3.1648
'chr_B_sg_C_Cara003g017860.1'	23.2579	1329.3670	1.3666	3.1236
'Ccan002g017880.1'	30.4651	903.0598	1.4838	2.9557
'chr_B_sg_E_Cara004g017990.1'	18.5378	1068.6318	1.2681	3.0288
'Ceug044g019540.1'	43.5332	1083.4981	1.6388	3.0348
Mean	55.8834	1002.3083	1.4993	2.9896
Median	30.4651	903.0598	1.4838	2.9557

TABLE A.53: Terpene Group 18.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'cal_g012193.t1'	1903.4226	1230.2049	3.2795	3.0900
'cal_g012195.t1'	1950.3991	1253.8213	3.2901	3.0982
CRO_T103423	1013.2515	858.9682	3.0057	2.9340
'Gese_snap_masked-scaffold_382-processed-gene-0.30-mRNA-1'	485.9870	485.9870	2.6866	2.6866
Mean	1338.2651	957.2453	3.0655	2.9522
Median	1458.3371	1044.5865	3.1426	3.0120

TABLE A.54: Terpene Group 19.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’Ceug027g022600.1’	1.5917	932.0179	0.2018	2.9694
’chr_J_sg_E_Cara020g014200.1’	1.5462	931.9953	0.1893	2.9694
’chr_J_sg_C_Cara019g013920.1’	186.3752	745.6809	2.2704	2.8726
’Ceug083g015360.1’	737.9400	1019.2935	2.8680	3.0083
’chr_G_sg_E_Cara014g022980.1’	11.0815	547.0790	1.0446	2.7381
’Ceug046g025900.1’	16.9117	549.9884	1.2282	2.7404
’Cara1063g005000.1’	11.8139	547.0443	1.0724	2.7380
’Ccan007g023000.1’	3.8563	810.7896	0.5862	2.9089
’Ccan007g022990.1’	347.7768	897.0664	2.5413	2.9528
’chr_G_sg_C_Cara013g021180.1’	526.6626	859.3972	2.7215	2.9342
Mean	184.5556	784.0352	1.4724	2.8832
Median	14.3628	835.0934	1.1503	2.9216

TABLE A.55: Terpene Group 20.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’cal_g013332.t1’	630.9539	358.0350	2.8000	2.5539
CRO_T118957	541.0000	313.0377	2.7332	2.4956
’Ccan001g031780.1’	8.0883	59.2720	0.9079	1.7728
’chr_A_sg_C_Cara001g026780.1’	6.3289	58.6854	0.8013	1.7685
’chr_A_sg_E_Cara002g028370.1’	16.5958	62.1028	1.2200	1.7931
’Ceug005g037840.1’	77.3598	82.3653	1.8885	1.9157
’Gese_maker-scaffold_175-snap-gene-2.73-mRNA-1’	371.3616	256.8157	2.5698	2.4096
’Solyc06g084240.1.1’	557.7084	368.6358	2.7464	2.5666
’Ipni_XP_019161709.1’	489.5197	327.7135	2.6898	2.5155
’Sein_XP_011079864.2’	530.1436	307.8825	2.7244	2.4884
’Gese_snap_masked-scaffold_934-processed-gene-0.3-mRNA-1’	2998.4830	1544.1366	3.4769	3.1887
’Utgi_unitig_26.g10150.t1’	1412.4604	749.2619	3.1500	2.8746
’Sein_XP_011082563.1’	567.6874	326.4955	2.7541	2.5139
’Oleu_XP_022864719.1’	101.4823	94.8368	2.0064	1.9770
’Oleu_XP_022898167.1’	109.6387	99.7305	2.0400	1.9988
GWHPAAAL023534	60.4824	67.4507	1.7816	1.8290
GWHPAAAL021059	73.3848	79.6472	1.8656	1.9012
GWHPAAAL014407	1.1588	60.2476	0.0640	1.7799
GWHPAAAL017992	68.3232	80.0027	1.8346	1.9031
GWHPAAAL022404	27.7700	49.1830	1.4436	1.6918
’Cac_g013051.t1’	155.2036	137.6211	2.1909	2.1387
’Cac_g015431.t1’	40.7007	62.7843	1.6096	1.7979
’Cac_g015430.t1’	68.4728	76.6709	1.8355	1.8846
’Cac_g015429.t1’	144.6781	114.7750	2.1604	2.0598
’Cac_g015426.t1’	0.4828	42.6745	-0.3162	1.6302
’TEA000496.1’	190.6127	177.3937	2.2802	2.2489
Achn308741	56.6350	64.3356	1.7531	1.8085
Achn308721	220.4135	183.4457	2.3432	2.2635
CRO_T136682	659.2712	502.6695	2.8191	2.7013
’Aqcoe1G360500.1.p’	937.8706	573.1066	2.9721	2.7582
’Cac_g033277.t1’	5608.2003	3246.2897	3.7488	3.5114
’Rico_XP_002520733.1’	186.1822	160.8546	2.2699	2.2064
’Potri.002G052100.1’	153.3647	125.9662	2.1857	2.1003
’Potri.005G210300.1’	140.6955	118.3648	2.1483	2.0732
’Prupe.8G239900.1.p’	227.8587	192.1116	2.3577	2.2836
Thecc1EG004358t1	159.4671	140.8178	2.2027	2.1487
GSVIVT01001153001	147.8808	137.3782	2.1699	2.1379
GSVIVT01001155001	78.6115	79.6545	1.8955	1.9012
’Nenu_XP_010277558.1’	117.9532	110.8635	2.0717	2.0448
’Nenu_XP_019053166.1’	117.0358	110.1427	2.0683	2.0420
’Aqcoe5G057100.1.p’	61.2635	66.3224	1.7872	1.8217
’Aqcoe4G243000.1.p’	123.7154	115.3905	2.0924	2.0622
Achn321931	172.6566	172.6566	2.2372	2.2372
Mean	428.3525	280.4146	2.1019	2.1814
Median	144.6781	118.3648	2.1604	2.0732

TABLE A.56: Terpene Group 21.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
CRO_T131521	288.1625	195.8760	2.4596	2.2920
’cal_g002829.t1’	597.9211	268.4407	2.7766	2.4288
’cal_g002830.t1’	1216.7619	326.6915	3.0852	2.5141
’cal_g002828.t1’	660.6005	215.2134	2.8199	2.3329
’Gese_augustus_masked-scaffold_408-processed-gene-0.4-mRNA-1’	194.9900	133.1395	2.2900	2.1243
’Gese_snap_masked-scaffold_408-processed-gene-0.26-mRNA-1’	992.4821	532.0492	2.9967	2.7260
Mean	658.4864	278.5684	2.7380	2.4030
Median	629.2608	241.8270	2.7983	2.3809

TABLE A.57: Terpene Group 22.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’TEA004822.1’	4.3480	4.3480	0.6383	0.6383
’TEA007191.1’	0.1796	1.7786	-0.7456	0.2501
’TEA002411.1’	0.000216	2.2518	-3.6653	0.3525
’TEA016700.1’	0.000216	2.2518	-3.6653	0.3525
Mean	1.1320	2.6575	-1.8595	0.3984
Median	0.0899	2.2518	-2.2054	0.3525

TABLE A.58: Terpene Group 23.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
Thecc1EG030063t1	119.5289	62.3318	2.0775	1.7947
’Prupe.4G030400.1.p’	115.9892	42.0867	2.0644	1.6241
’Prupe.4G030300.1.p’	172.5726	60.9484	2.2370	1.7850
Thecc1EG030061t1	17.3866	20.5853	1.2402	1.3136
Thecc1EG030060t1	38.6053	28.7245	1.5866	1.4583
Thecc1EG030064t1	61.2904	36.2864	1.7874	1.5597
’AT1G61680.1’	137.5657	99.6388	2.1385	1.9984
’Potri.011G032300.1’	82.9989	59.3131	1.9191	1.7732
’Rico_XP_025012989.1’	81.6105	58.4800	1.9117	1.7670
Mean	91.9498	52.0439	1.8847	1.6749
Median	82.9989	58.4800	1.9191	1.7670

TABLE A.59: Terpene Group 24.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’chr_E_sg_E_Cara010g015270.1’	14.6018	16.5684	1.1644	1.2193
’Ceug007g020060.1’	23.3236	20.1318	1.3678	1.3039
’Ceug007g019530.1’	10.4795	15.8495	1.0203	1.2000
’Ccan005g017150.1’	16.5221	17.5287	1.2181	1.2437
’chr_E_sg_C_Cara009g014750.1’	26.1296	22.3328	1.4171	1.3489
’Ccan005g017160.1’	13.0958	9.8480	1.1171	0.9933
’chr_E_sg_C_Cara009g014760.1’	0.2064	3.4030	-0.6854	0.5319
Mean	14.9084	15.0946	0.9456	1.1202
Median	14.6018	16.5684	1.1644	1.2193

TABLE A.60: Terpene Group 25.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’AT1G70080.1’	200.9848	166.4663	2.3032	2.2213
’AT1G33750.1’	489.1065	203.3968	2.6894	2.3083
’AT3G29410.1’	950.1466	223.0008	2.9778	2.3483
’AT3G32030.1’	788.7054	205.0344	2.8969	2.3118
’AT3G14490.1’	152.1868	138.6925	2.1824	2.1421
’AT1G31950.1’	113.9866	127.0470	2.0569	2.1040
’AT3G14540.1’	146.8413	134.2974	2.1668	2.1281
’AT3G14520.1’	112.3209	128.8459	2.0505	2.1101
Thecc1EG041391t1	120.3209	91.0692	2.0803	1.9594
Thecc1EG032807t1	97.6758	79.7465	1.9898	1.9017
Mean	317.2276	149.7597	2.3394	2.1535
Median	149.5141	136.4950	2.1746	2.1351

TABLE A.61: Terpene Group 26.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’cal_g014841.t1’	870.4984	457.0694	2.9398	2.6600
CRO_T137624	1429.8729	391.1017	3.1553	2.5923
’cal_g011304.t1’	1110.9466	311.0515	3.0457	2.4928
GWHPAAAL017694	0.0006	32.3085	-3.2460	1.5093
GWHPAAAL020731	0.0006	32.3085	-3.2460	1.5093
CRO_T111308	285.4693	285.4693	2.4556	2.4556
Mean	616.1314	251.5515	0.8507	2.2032
Median	577.9839	298.2604	2.6977	2.4742

TABLE A.62: Terpene Group 27.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’cal_g019208.t1’	429.7563	429.7563	2.6332	2.6332
’cal_g011259.t1’	912.5623	482.9670	2.9603	2.6839
CRO_T123707	930.3323	491.8604	2.9686	2.6918
’Gese_snap_masked-scaffold_423-processed-gene-0.14-mRNA-1’	728.4407	503.2721	2.8624	2.7018
Mean	750.2729	476.9640	2.8561	2.6777
Median	820.5015	487.4137	2.9113	2.6879

TABLE A.63: Terpene Group 28.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
’chr_H_sg_E_Cara016g014450.1’	33.46564753	241.7971044	1.524599233	2.383451096
’Ceug021g020500.1’	14.30889341	232.2169705	1.155606048	2.365893955
’Ccan008g017480.1’	6.133103628	228.1383542	0.7876803028	2.358198304
’chr_H_sg_C_Cara015g014260.1’	0.4458380102	225.2943282	-0.3508229083	2.352750258
’Ceug021g020640.1’	0.6121620539	257.5349007	-0.2131335945	2.410836092
’Ccan008g017540.1’	85.73597193	294.0141579	1.933163076	2.468368244
’Ccan008g017550.1’	216.6404189	310.0394403	2.335739487	2.491416944
’chr_H_sg_C_Cara015g014270.1’	51.36958319	362.5435064	1.710706042	2.559360131
’Ccan008g017490.1’	35.87526605	359.1379539	1.554795131	2.555261304
’chr_H_sg_E_Cara016g014460.1’	133.0719031	380.3287286	2.124086368	2.580159132
’Cara624g005000.1’	1097.796781	507.892113	3.040521953	2.705771469
’chr_H_sg_E_Cara016g014480.1’	2079.511281	595.3426947	3.317961281	2.774767029
’Ceug021g020510.1’	2.37041213	379.2922085	0.3748238608	2.578973921
’chr_H_sg_E_Cara016g012420.1’	354.7383743	433.5141698	2.549908171	2.637003297
’Ccan008g015810.1’	65.95004165	383.0241025	1.819215074	2.583226104
’Ceug021g019100.1’	56.29829845	333.2827554	1.750495269	2.522812843
’Ceug021g020620.1’	21.82478607	375.4418999	1.338949995	2.574542739
’Ceug021g019070.1’	51.27398352	380.5815239	1.709897059	2.580447701
’Ccan008g017500.1’	44.08568343	379.3316354	1.644297578	2.579019063
’chr_H_sg_C_Cara015g014360.1’	24.74067304	375.9556671	1.39341151	2.575136636
’Cara2873g005020.1’	256.6997317	407.6070603	2.409425415	2.610241698
’chr_H_sg_C_Cara015g014280.1’	375.6225544	433.7865386	2.574751662	2.637276071
’Cara2872g005000.1’	499.0185632	467.79292	2.698116701	2.670053645
’Ceug021g020570.1’	2.102502666	371.7905586	0.3227365552	2.570298357
’Cara625g005030.1’	206.1922792	425.6881263	2.314272399	2.629091537
’chr_H_sg_E_Cara016g013360.1’	5515.488151	959.1329494	3.741583956	2.981878811
’chr_H_sg_E_Cara016g014550.1’	179.5272949	402.8392118	2.254130487	2.605131738
’chr_H_sg_E_Cara016g014610.1’	801.9044121	511.7487404	2.904122603	2.709056782
’Cara6411g005010.1’	260.4589559	376.6283201	2.415739295	2.575912973
’Ccan008g017570.1’	92.48608368	356.9213319	1.96607639	2.552572505
’Ceug021g020720.1’	201.0515876	423.8284681	2.303307507	2.627190124
’chr_H_sg_E_Cara016g014720.1’	40.85440732	406.8692557	1.611238915	2.609454874
’Ceug021g023940.1’	39.85198578	372.1612512	1.600449967	2.570731153
’Ceug007g011810.1’	80.70324483	297.6790639	1.906890997	2.473748291
’Ceug005g012950.1’	96.44698496	328.0043163	1.984288656	2.515879559
’chr_H_sg_E_Cara016g014740.1’	309.2351851	401.5089778	2.490288902	2.603695261
’chr_H_sg_C_Cara015g014350.1’	199.9178345	386.8415121	2.300851539	2.587533072
’chr_H_sg_E_Cara016g014530.1’	3.896150345	404.3663615	0.5906357071	2.60677502
’Ceug021g020560.1’	0.005012588972	425.4280808	-2.299937906	2.628826153
’Ceug021g019080.1’	0.005012588972	425.4280808	-2.299937906	2.628826153
’Cara2872g005010.1’	102.0478754	386.1281667	2.008803967	2.586731483
’Cara625g005000.1’	0.6603574104	268.9918348	-0.1802209443	2.429739097
’chr_H_sg_E_Cara016g014650.1’	97.25990675	361.0166519	1.987933849	2.557527234
’Cara2873g005030.1’	165.1357876	378.1687179	2.217841202	2.577685601
’Ceug021g020610.1’	8.231485934	272.0618626	0.9154782404	2.434667667
’chr_H_sg_E_Cara016g014590.1’	7.766161549	196.1494449	0.8902064204	2.292587083
’Ccan005g013690.1’	30.31600635	283.7513774	1.481671989	2.452937978
’chr_H_sg_E_Cara016g014680.1’	1.356610253	339.1586368	0.1324550951	2.530402881
’Ceug021g020630.1’	264.4065883	404.2379453	2.422272272	2.606637078
’Ceug007g011820.1’	10.95801097	200.9229431	1.039731731	2.303029531
’Ceug007g016180.1’	18.8032029	69.21393115	1.274231832	1.840193517
’Ceug097g018420.1’	662.397736	255.3433722	2.821118839	2.40712459
’Ceug044g040490.1’	2.003663371	102.8077962	0.301824759	2.01202605
’chr_B_sg_E_Cara004g036890.1’	2.003663371	102.8077962	0.301824759	2.01202605
’chr_H_sg_C_Cara015g014510.1’	0.5008963877	71.93477962	-0.3002521003	1.856938917
’Ccan003g023860.1’	11.90032781	83.73214445	1.075558925	1.922892214
’chr_C_sg_C_Cara005g021700.1’	50.88476	96.72388796	1.70658773	1.985533745
’Ccan003g023850.1’	548.6528617	334.2312345	2.739297649	2.524047033
’Ceug097g018410.1’	393.0891699	236.858752	2.594491079	2.374489437
’Ceug044g040730.1’	132.2928406	121.6385091	2.121536342	2.085071088
’Ccan002g037580.1’	56.8918142	109.705575	1.755049783	2.040228698
’Ceug044g040500.1’	18.75951149	92.11096093	1.273221525	1.964311313
’chr_B_sg_E_Cara004g036900.1’	91.66299166	113.4809675	1.962194027	2.05492303
’Cara2583g005000.1’	1034.948334	233.1370698	3.01491867	2.367611334
’chr_B_sg_C_Cara003g035560.1’	2.362853366	105.093176	0.373436771	2.021574517
Mean	265.4309	320.6179	1.5573	2.4503
Median	56.2983	359.1380	1.7505	2.5553

TABLE A.64: Terpene Group 29.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'Ceug097g029870.1'	18.0260	29.3931	1.2559	1.4682
'Ceug027g011480.1'	0.3594	27.2929	-0.4445	1.4360
'chr_D_sg_E_Cara008g023440.1'	14.8197	32.1134	1.1708	1.5067
'chr_D_sg_C_Cara007g023000.1'	3.7659	22.2630	0.5759	1.3476
'Ccan004g024620.1'	11.2130	25.9867	1.0497	1.4148
'chr_D_sg_C_Cara007g022990.1'	10.7855	25.7731	1.0328	1.4112
'Ccan004g024610.1'	5.3453	23.0529	0.7280	1.3627
'Ceug027g011470.1'	47.6707	44.2162	1.6783	1.6456
'chr_D_sg_E_Cara008g023430.1'	6.5703	29.3634	0.8176	1.4678
'Ceug097g029860.1'	24.1150	35.2121	1.3823	1.5467
'Ceug097g030500.1'	0.2510	45.2552	-0.6003	1.6557
'Ccan004g025070.1'	0.2246	45.2420	-0.6486	1.6555
'Cara2018g005000.1'	23.9359	46.0437	1.3790	1.6632
'chr_D_sg_E_Cara008g023860.1'	30.8423	45.6979	1.4891	1.6599
'chr_D_sg_E_Cara008g023230.1'	0.2576	45.2581	-0.5890	1.6557
'Ceug034g031220.1'	5.0243	47.6419	0.7011	1.6780
'Ccan004g024410.1'	0.3762	54.2907	-0.4245	1.7347
'Ceug027g011330.1'	290.8517	140.4131	2.4637	2.1474
'chr_D_sg_E_Cara008g023250.1'	5919.7031	936.1212	3.7723	2.9713
'Ccan004g024420.1'	457.9068	143.6171	2.6608	2.1572
'chr_D_sg_E_Cara008g023280.1'	12.6397	66.3783	1.1017	1.8220
'Ceug027g011360.1'	21.1649	69.0023	1.3256	1.8389
'Ceug034g031250.1'	4.9542	50.1163	0.6950	1.7000
'Ccan004g024470.1'	3.2616	38.0597	0.5134	1.5805
'chr_D_sg_C_Cara007g022820.1'	0.1630	36.2003	-0.7879	1.5587
Mean	276.5691	84.1602	0.8919	1.6834
Median	10.7855	45.2420	1.0328	1.6555

TABLE A.65: Terpene Group 30.

Gene name	Quotient	Gene-to-node	Log Quotient	Log Gene-to-node
'Ccan002g037410.1'	0.4281	140.7498	-0.3684	2.1484
'chr_B_sg_C_Cara003g035710.1'	0.4281	140.7498	-0.3684	2.1484
'chr_B_sg_E_Cara004g036770.1'	49.7928	203.9262	1.6972	2.3095
'Ceug044g040350.1'	160.8195	240.9449	2.2063	2.3819
'Ccan002g037420.1'	0.9564	187.6609	-0.0193	2.2734
'chr_B_sg_C_Cara003g035700.1'	0.8558	187.6273	-0.0676	2.2733
'Ceug044g040370.1'	0.9564	187.6579	-0.0194	2.2734
'Ceug044g049510.1'	0.8558	187.6243	-0.0676	2.2733
'Ceug044g040420.1'	3903.1578	1194.2437	3.5914	3.0771
'Ceug097g018580.1'	0.0017	240.7510	-2.7761	2.3816
'Ceug044g049530.1'	0.0017	240.7510	-2.7761	2.3816
'chr_B_sg_E_Cara004g036790.1'	88.3805	232.7861	1.9464	2.3670
'Ceug097g018570.1'	202.7435	269.7989	2.3069	2.4310
'Ceug044g040410.1'	45.3142	247.2250	1.6562	2.3931
'Ccan005g024220.1'	43.6912	138.6248	1.6404	2.1418
'Ceug007g026810.1'	201.4296	254.4783	2.3041	2.4057
'Ceug007g026830.1'	189.5143	250.5044	2.2776	2.3988
'Ccan005g024210.1'	97.2203	235.0363	1.9878	2.3711
'chr_E_sg_C_Cara009g021660.1'	8.0302	212.7211	0.9047	2.3278
'Ccan001g021400.1'	1.2831	211.0559	0.1083	2.3244
'chr_A_sg_E_Cara002g014490.1'	1.3585	211.0748	0.1331	2.3244
'chr_H_sg_C_Cara015g014430.1'	6.9004	84.3596	0.8389	1.9261
'Ceug005g012980.1'	15.3178	132.4352	1.1852	2.1220
'Ceug021g020830.1'	2111.2700	552.8669	3.3245	2.7426
'chr_H_sg_C_Cara015g014380.1'	289.3539	187.3511	2.4614	2.2727
'Ccan008g017590.1'	214.7785	172.4232	2.3320	2.2366
'Ccan008g017640.1'	18.1905	100.2628	1.2598	2.0011
Mean	283.4456	246.1367	1.0259	2.3225
Median	18.1905	211.0559	1.2598	2.3244

TABLE A.66: Terpene Group 31.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
'cal_g007979.t1'	974.7773	605.7526	2.9889	2.7823
CRO_T125942	808.3622	522.4551	2.9076	2.7180
'Gese_snap_masked-scaffold_203-processed-gene-2.47-mRNA-1'	628.1389	379.4830	2.7981	2.5792
'Gese_augustus_masked-scaffold_203-processed-gene-2.19-mRNA-1'	587.0349	283.0914	2.7687	2.4519
'Gese_maker-scaffold_2048-snap-gene-0.3-mRNA-1'	938.9688	293.1160	2.9727	2.4670
'Gese_maker-scaffold_32-augustus-gene-8.68-mRNA-1'	469.1140	198.5871	2.6713	2.2980
CRO_T127456	738.8717	536.2040	2.8686	2.7293
Mean	735.0383	402.6699	2.8537	2.5751
Median	738.8717	379.4830	2.8686	2.5792

TABLE A.67: Terpene Group 32.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
'Ipni_XP_019194074.1'	240.5271	240.5271	2.3812	2.3812
'Solyc06g060180.1.1'	502.6652	364.7542	2.7013	2.5620
'Solyc06g060010.2.1'	869.2638	548.1696	2.9392	2.7389
'Solyc06g059930.2.1'	316.9213	262.8087	2.5010	2.4196
'Solyc06g059910.2.1'	1.4616	169.3875	0.1648	2.2289
'Gese_snap_masked-scaffold_872-processed-gene-0.5-mRNA-1'	10707.0497	2892.3333	4.0297	3.4612
'Solyc09g092470.2.1'	597.9203	438.8746	2.7766	2.6423
CRO_T140431	694.8317	577.8254	2.8419	2.7618
Mean	1741.3301	686.8351	2.5419	2.6495
Median	550.2927	401.8144	2.7390	2.6022

TABLE A.68: Terpene Group 33.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
'TEA022784.1'	151.5373	105.0179	2.1805	2.0213
'TEA014184.1'	81.4280	69.9630	1.9108	1.8449
'Gese_augustus_masked-scaffold_286-processed-gene-1.1-mRNA-1'	113.5130	167.6632	2.0550	2.2244
'chr_B_sg_E_Cara004g017020.1'	362.2048	292.1488	2.5590	2.4656
'evm_27.model.AmTr_v1.0_scaffold00115.23'	478.7523	350.4938	2.6801	2.5447
'AT4G21320.1'	545.9369	330.1509	2.7371	2.5187
GSVIVT01002901001	191.8646	211.8917	2.2830	2.3261
GWHPAAAL015545	118.4730	144.3141	2.0736	2.1593
'TEA021924.1'	126.5057	145.5535	2.1021	2.1630
Mean	241.1351	201.9108	2.2868	2.2520
Median	151.5373	167.6632	2.1805	2.2244

TABLE A.69: Terpene Group 34.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
'Aqcoe4G298300.1.p'	15.8149	9.4673	1.1991	0.9762
'Aqcoe5G394600.1.p'	10.8044	5.6813	1.0336	0.7544
'Aqcoe4G237300.1.p'	16.8459	7.6951	1.2265	0.8862
'Aqcoe4G236500.1.p'	20.1556	11.7302	1.3044	1.0693
'Aqcoe5G393700.1.p'	40.2463	8.5823	1.6047	0.9336
'Aqcoe4G067000.1.p'	41.3454	8.7393	1.6164	0.9415
'Aqcoe4G128100.1.p'	49.5909	9.9172	1.6954	0.9964
'Aqcoe7G284900.1.p'	153.4101	14.8519	2.1859	1.1718
'Aqcoe4G125200.1.p'	63.3910	7.9271	1.8020	0.8991
'Aqcoe4G248400.1.p'	19.9074	8.8391	1.2990	0.9464
'Aqcoe4G127600.1.p'	11.0021	6.3838	1.0415	0.8051
'Aqcoe4G273200.1.p'	13.2902	7.8437	1.1235	0.8945
Mean	37.9837	8.9715	1.4277	0.9395
Median	20.0315	8.6608	1.3017	0.9375

TABLE A.70: Terpene Group 35.

Gene name	**Quotient**	**Gene-to-node**	**Log Quotient**	**Log Gene-to-node**
'Aqcoe7G440300.1.p'	3.58777795	3.58777795	0.5548	0.5548
'Aqcoe4G253500.1.p'	0.1402582595	1.176694497	-0.8531	0.0707
'Aqcoe4G254000.1.p'	0.2186211243	1.215875929	-0.6603	0.0849
'Aqcoe4G034600.1.p'	1.898181673	2.055656197	0.2783	0.3130
'Aqcoe4G129400.1.p'	0.0174805446	1.664218183	-1.7574	0.2212
'Aqcoe4G129600.1.p'	0.1786984103	1.70452265	-0.7479	0.2316
'Aqcoe7G261300.1.p'	0.9618276089	1.900304877	-0.0169	0.2788
'Aqcoe7G261500.1.p'	6.021471503	3.165216178	0.7797	0.5004
Mean	1.6280	2.0588	-0.3028	0.2819
Median	0.5902	1.8024	-0.3386	0.2552

Bibliography

Albert, Victor A. et al. (2013). “The Amborella Genome and the Evolution of Flowering Plants”. In: *Science* 342.6165.

Ashihara, Hiroshi and Alan Crozier (2001). “Caffeine: a well known but little mentioned compound in plant science”. In: *Trends in Plant Science* 6.9, pp. 407 –413.

Bolker, B. et al. (2011). “phylobase: Base Package for Phylogenetic Structures and Comparative Data.” In: *R Hackathon*.

Bolton, Sanford and Gary Null (1981). “Caffeine: Psychological effects, use and abuse”. In: *Journal of Orthomolecular Psychiatry* 10.3, pp. 202–211.

De Kochko, Alexandre, Dominique Crouzillat, and Arabica Consortium (Oct. 2015). “Aims and goals of the Arabica Coffee Genome Consortium (ACGC)”. In: 12th Solanaceae Conference.

Denoeud, France et al. (2014). “The coffee genome provides insight into the convergent evolution of caffeine biosynthesis”. In: *Science* 345.6201, pp. 1181–1184.

Du, Qingzhang et al. (2015). “Populus endo--1,4-glucanases gene family: genomic organization, phylogenetic analysis, expression profiles and association mapping”. In: *Planta* 241.6, pp. 1417–1434.

Filiault, Danièle L et al. (2018). “The Aquilegia genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history”. In: *eLife* 7, e36426.

Franke, Jakob et al. (2019). “Gene Discovery in Gelsemium Highlights Conserved Gene Clusters in Monoterpene Indole Alkaloid Biosynthesis”. In: *ChemBioChem* 20.1, pp. 83–87.

Hoopes, Genevieve M. et al. (2018). “Genome Assembly and Annotation of the Medicinal Plant Calotropis gigantea, a Producer of Anticancer and Antimalarial Cardenolides”. In: *G3: Genes|Genomes|Genetics* 8.2, p. 385.

Hoshino, Atsushi et al. (2016). “Genome sequence and analysis of the Japanese morning glory Ipomoea nil”. In: *Nature Communications* 7.1, p. 13295.

Huang, Ruiqi et al. (2016). “Convergent evolution of caffeine in plants by co-option of exapted ancestral enzymes”. In: *Proceedings of the National Academy of Sciences* 113.38, pp. 10613–10618.

Huang, Shengxiong et al. (2013). “Draft genome of the kiwifruit Actinidia chinensis”. In: *Nature Communications* 4.1, p. 2640.

Jaillon, O. et al. (2007). “The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla”. In: *Nature* 449.7161, pp. 463–467.

Kellner, Franziska et al. (2015). “Genome-guided investigation of plant natural product biosynthesis”. In: *The Plant Journal* 82.4, pp. 680–692.

Klopfstein, Seraina, Tim Massingham, and Nick Goldman (2017). “More on the Best Evolutionary Rate for Phylogenetic Analysis”. In: *Systematic Biology* 66.5, pp. 769–785.

Kumar, S. et al. (2017). “TimeTree: a resource for timelines, timetrees, and divergence times”. In: *Molecular Biology and Evolution* 34, pp. 1812–1819. DOI: `DOI:10.1093/molbev/msx116`.

Lamesch, Philippe et al. (2011). “The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools”. In: *Nucleic Acids Research* 40.D1, pp. D1202–D1210.

Lan, T. et al. (2017). “Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome”. In: *Proceedings of the National Academy of Sciences* 114.22, E4435.

Lenton, Timothy M. et al. (2016). “Earliest land plants created modern levels of atmospheric oxygen”. In: 113.35, pp. 9704–9709. DOI: `10.1073/pnas.1604787113`.

Lynch, Michael et al. (2016). “Genetic drift, selection and the evolution of the mutation rate”. In: *Nature Reviews Genetics* 17.11, pp. 704–714.

Ming, Ray et al. (2013). “Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.)” In: *Genome Biology* 14.5, R41.

Mirarab, Siavash et al. (2015). “PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences”. In: *Journal of computational biology : a journal of computational molecular cell biology* 22.5, pp. 377–386.

Motamayor, Juan C. et al. (2013). “The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color”. In: *Genome Biology* 14.6, r53.

Ogawa, Mikihiro et al. (2001). “7-Methylxanthine Methyltransferase of Coffee Plants”. In: *Journal of Biological Chemistry* 276, pp. 8213–8218.

Ohta, Tomoko (1994). “Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory”. In: *Journal of Molecular Evolution* 40.1, pp. 56–63.

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer (2004). “APE: Analyses of Phylogenetics and Evolution in R language”. In: *Bioinformatics* 20.2, pp. 289–290.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin (2010). “FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments”. In: *PLOS ONE* 5.3, pp. 1–10.

Reis, Mario dos and Ziheng Yang (Sept. 2013). “Why Do More Divergent Sequences Produce Smaller Nonsynonymous/Synonymous Rate Ratios in Pairwise Sequence Comparisons?” In: *Genetics* 195.1, pp. 195–204.

Rivarola, Maximo et al. (2011). “Castor Bean Organelle Genome Sequencing and Worldwide Genetic Diversity Analysis”. In: *PLOS ONE* 6.7, e21743.

Robinson, David (2014). “broom: An R package for converting statistical analysis objects into Tidy Data Frames”. In: *arXiv*.

Sanderson, Michael J. (2002). “Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach”. In: *Molecular Biology and Evolution* 19.1, pp. 101–109.

Santos Muñoz, Daniella, Eric Lam, and David Sankoff (2019). “Evolutionary Rate Change and the Transformation from Additive to Ultrametric: Modal Similarity of Orthologs in Fish and Flower Phylogenomics”. In: *Bioinformatics and Phylogenetics: Seminal Contributions of Bernard Moret*. Ed. by Tandy Warnow. Vol. 29. Springer International Publishing. Chap. 8, pp. 175–192.

Sato, Shusei et al. (2012). “The tomato genome sequence provides insights into fleshy fruit evolution”. In: *Nature* 485.7400, pp. 635–641.

Schliep, K.P. (2011). “phangorn: phylogenetic analysis in R”. In: *Bioinformatics* 27.4, pp. 592–593. DOI: `10.1093/bioinformatics/btq706`.

Unver, Turgay et al. (2017). “Genome of wild olive and the evolution of oil biosynthesis”. In: *Proceedings of the National Academy of Sciences* 114.44, E9413.

Verde, Ignazio et al. (2013). “The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution”. In: *Nature Genetics* 45.5, pp. 487–494.

Wang, Linhai et al. (2016). “Updated sesame genome assembly and fine mapping of plant height and seed coat color QTLs using a new high-density genetic map”. In: *BMC Genomics* 17.1, p. 31.

Wolf, Jochen B. W. et al. (Aug. 2009). “Nonlinear Dynamics of Nonsynonymous (dN) and Synonymous (dS) Substitution Rates Affects Inference of Selection”. In: *Genome Biology and Evolution* 1, pp. 308–319.

Xia, En-Hua et al. (2019). “Tea Plant Information Archive: a comprehensive genomics and bioinformatics platform for tea plant”. In: *Plant Biotechnology Journal* 17.10, pp. 1938–1953.

Yang, Z (Oct. 1997). “PAML: a program package for phylogenetic analysis by maximum likelihood.” In: *Comput Appl Biosci* 13.5, pp. 555–6.

Yang, Z, R Nielsen, and M Hasegawa (Dec. 1998). “Models of amino acid substitution and applications to mitochondrial protein evolution.” In: *Molecular Biology and Evolution* 15.12, pp. 1600–1611.

Zhao, Dongyan et al. (2017). “De novo genome assembly of Camptotheca acuminata, a natural source of the anti-cancer compound camptothecin”. In: *GigaScience* 6.9.