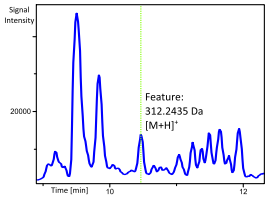
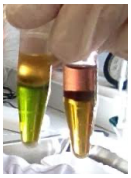
Supplementary figures



Spectral data generation using LC-qToF-HRMS

Extraction in 50% acetonitrile

Leaf tissue samples



DNA isolation

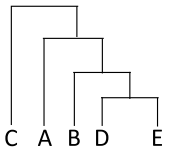
nuclear ribosomal DNA sequencing and alignment

Raw spectral data processing using: MZmine 2

	m/z	RT [min]	Signal intensity	
			Specimen A	Specimen B
Feature 1	312.2435	10.45	15000	0
Feature 2	276.2735	15.69	5000	50000
Feature 3	586.3546	25.23	30000	5000

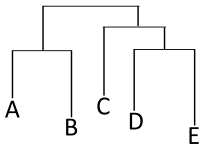
Export list of spectral features with associated MS^1 and MS^2 spectral information.

Metabolic cluster analysis



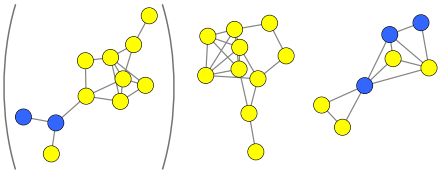
Displays chemical similarity among specimen

Phylogenetic analysis



Displays evolutionary relationships among specimen

Feature-based molecular networking (FBMN) at GNPS



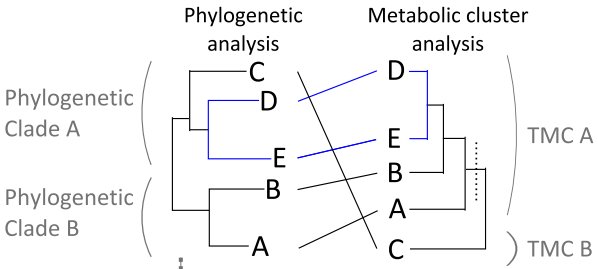
Chemical family:
Subset of nodes that share spectral similarity, indicating a cluster of structurally related metabolites.

- **Node**: Represents a spectral feature (=putative metabolite) with associated MS^1 and MS^2 spectral data.
- **Dereplicated node**: Structural annotation of node by spectral similarity to reference metabolite.
- **Edge**: Displays spectral similarity between nodes.

Additional networking analyses in GNPS

- Spectral database comparison using public, in house and *in silico* reference spectra.
- Network Annotation Propagation (NAP) predicts chemical classification.
- Unsupervised substructure discovery by MS2LDA reveals substructural motifs.

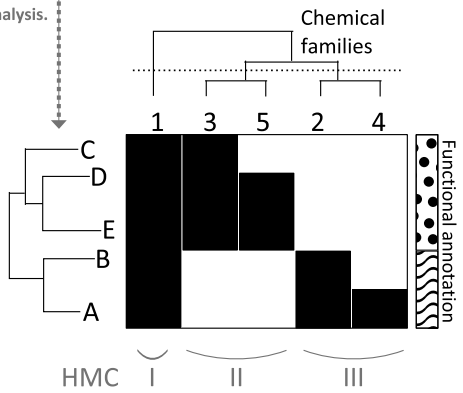
Tanglegram analysis



Tanglegram metabolic cluster (TMC): Represents a subset of samples with similar chemical profile.

Tanglegram-refined phylogeny is used for heatmap analysis.

Categorical heatmap analysis



Presence of chemical family based on at least one corresponding spectral feature present in the sample.

Absence of chemical family based on complete lack of corresponding spectral features.

Heatmap metabolic cluster (HMC): Represents a subset of chemical families with similar phylogenetic distribution.

Figure S1. Schematic workflow of the chemo-evolutionary analysis. 291 specimen involving species and subspecies of the tribe Myoporeae were investigated regarding their chemical and phylogenetic relationships. LC-qToF-HRMS was applied on crude leaf extracts to generate mass spectrometric data including MS^1 and MS^2. MZmine2 was used to isolate spectral features from the raw data and align those across all samples to output a feature table that contains the relatively quantified chemical composition of the dataset. Spectral information of those isolated features is then exported and integrated into the feature-based molecular networking (FBMN) pipeline at GNPS (https://gnps.ucsd.edu). The generated molecular network is composed of nodes and edges that represent individual spectral features and shared spectral similarity, respectively. Spectral similar nodes fall into clusters of varying size that are considered as a chemical family of structurally related metabolites. The molecular network provides the foundation for an exhaustive dereplication approach involving public, in house and *in silico* libraries of reference metabolites for spectral comparison. Dereplicated nodes are structurally annotated and utilized to achieve a prediction of chemical classification for each chemical family by using the Network Annotation Propagation (NAP) method. Using the unsupervised substructure discovery approach MS2LDA, substructural motifs can be predicted within the given spectral data and annotated throughout the molecular network. To compare the chemical profiles of all 291 specimen, the presence (at least one spectral feature present of a corresponding chemical family) and absence (no feature present in the sample) information of chemical families can be used. This approach yields a binary dataset that can be compiled by a metabolic cluster analysis into a dendrogram that displays chemical similarity among the specimen. In contrast, a phylogenetic analysis shows the evolutionary relationships among those specimen and is based on nuclear ribosomal DNA sequencing of leaf tissue from the same specimens. A tanglegram analysis directly compares the analyses by rearranging branches in both dendrograms to reveal similar clustering (i.e. specimen D and E) and thus chemo-evolutionary relationships. Groups of specimen that display an evolutionary lineage are referred to as phylogenetic clades, while chemical similarity is shared within tanglegram metabolic cluster (TMC). To extend the view on the chemo-evolutionary framework, a categorical heatmap analysis is used that clusters the presence/absence of individual chemical families according to their phylogenetic distribution alongside the tanglegram-refined phylogeny. Heatmap metabolic cluster (HMC) can be derived from this approach, which represent subsets of chemical families that share a similar phylogenetic signature. Additional functional annotations can be integrated to test metadata for correlation with the chemo-evolutionary framework that is displayed by the heatmap.

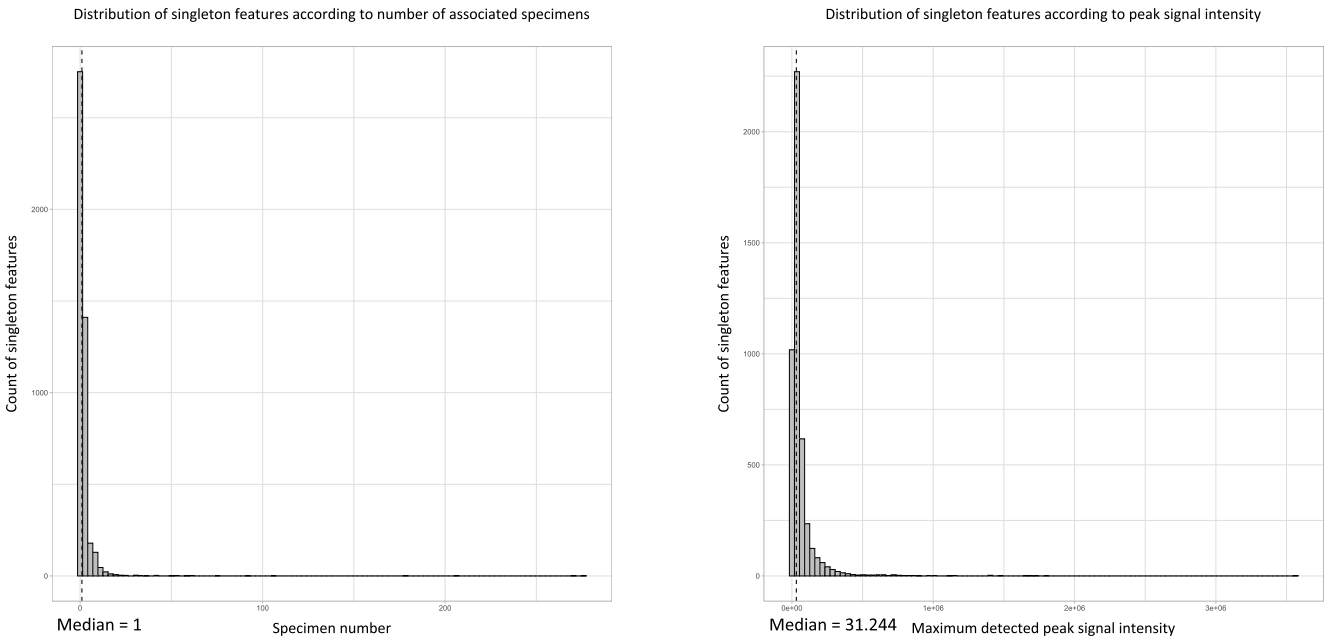
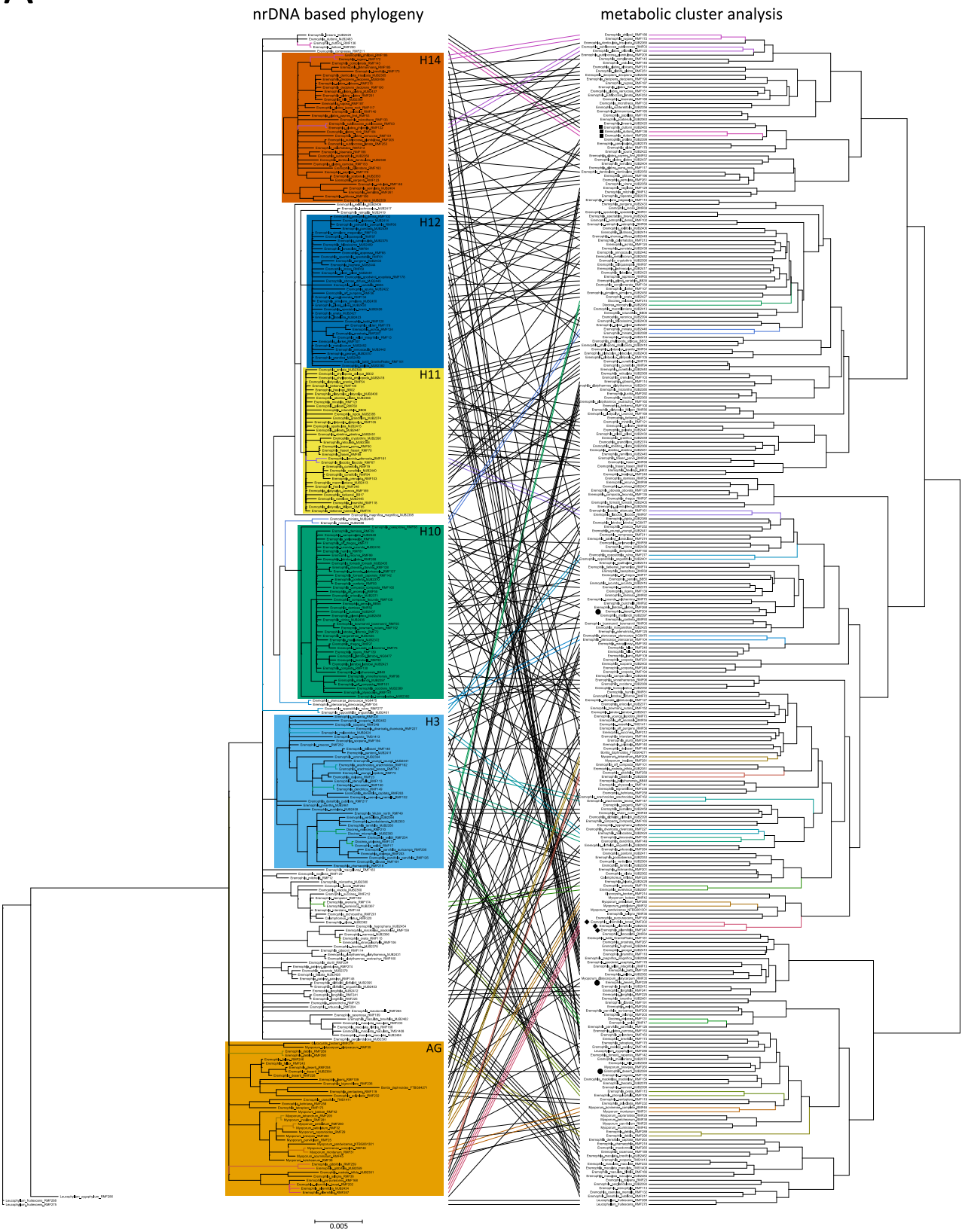


Figure S2. Specimen and peak signal intensity distribution of singleton features. Distribution graphs showing the number of singleton features according to number of associated specimens and peak signal intensity (detected maximum value for each feature among all specimens). Calculated medians are highlighted by a dotted vertical line and the actual value displayed below each respective graph.

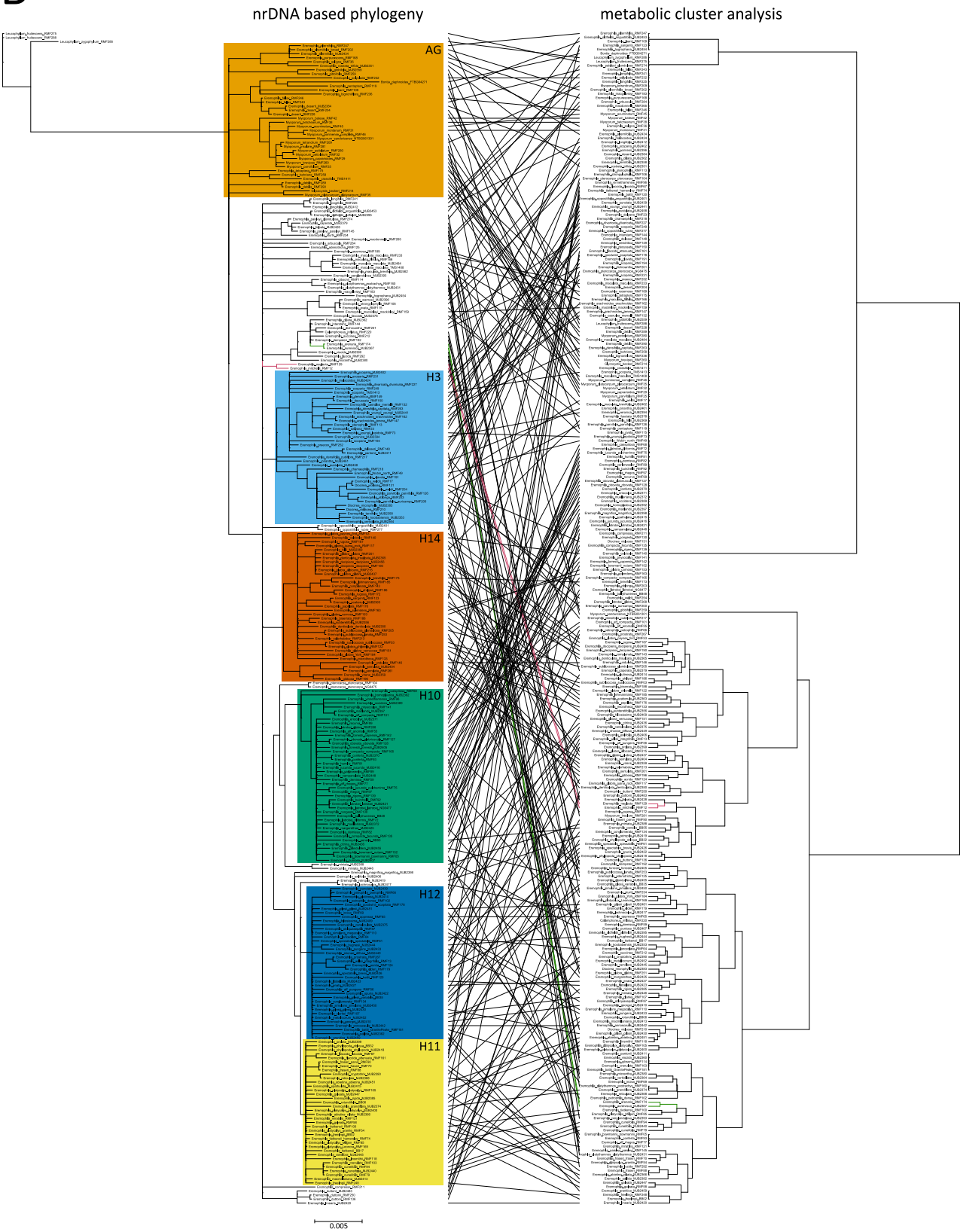
A

Tanglegram based on 10696 individual spectral features (after normalisation)



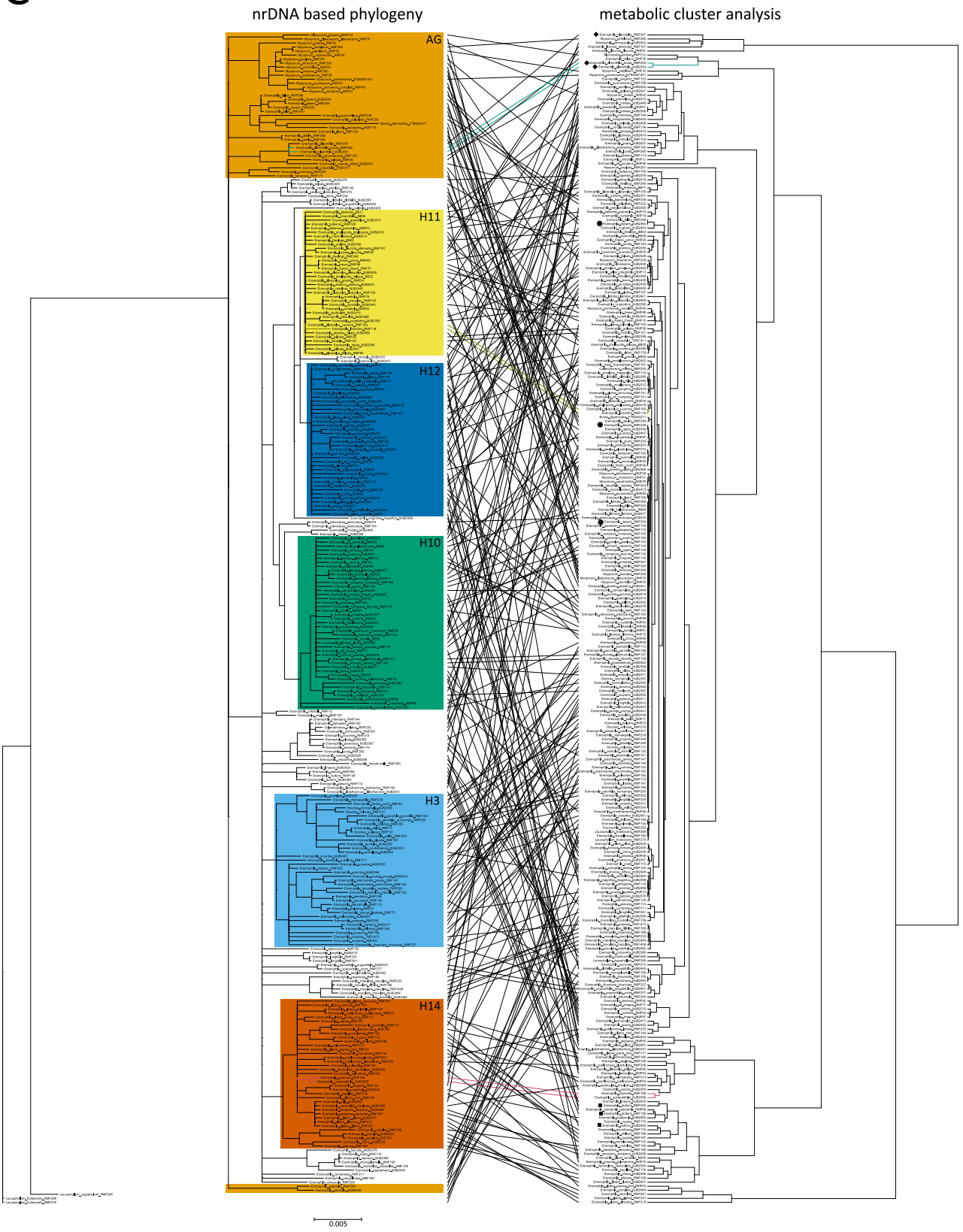
B

Tanglegram based on serrulatane- and viscidane diterpenoid related HMCs

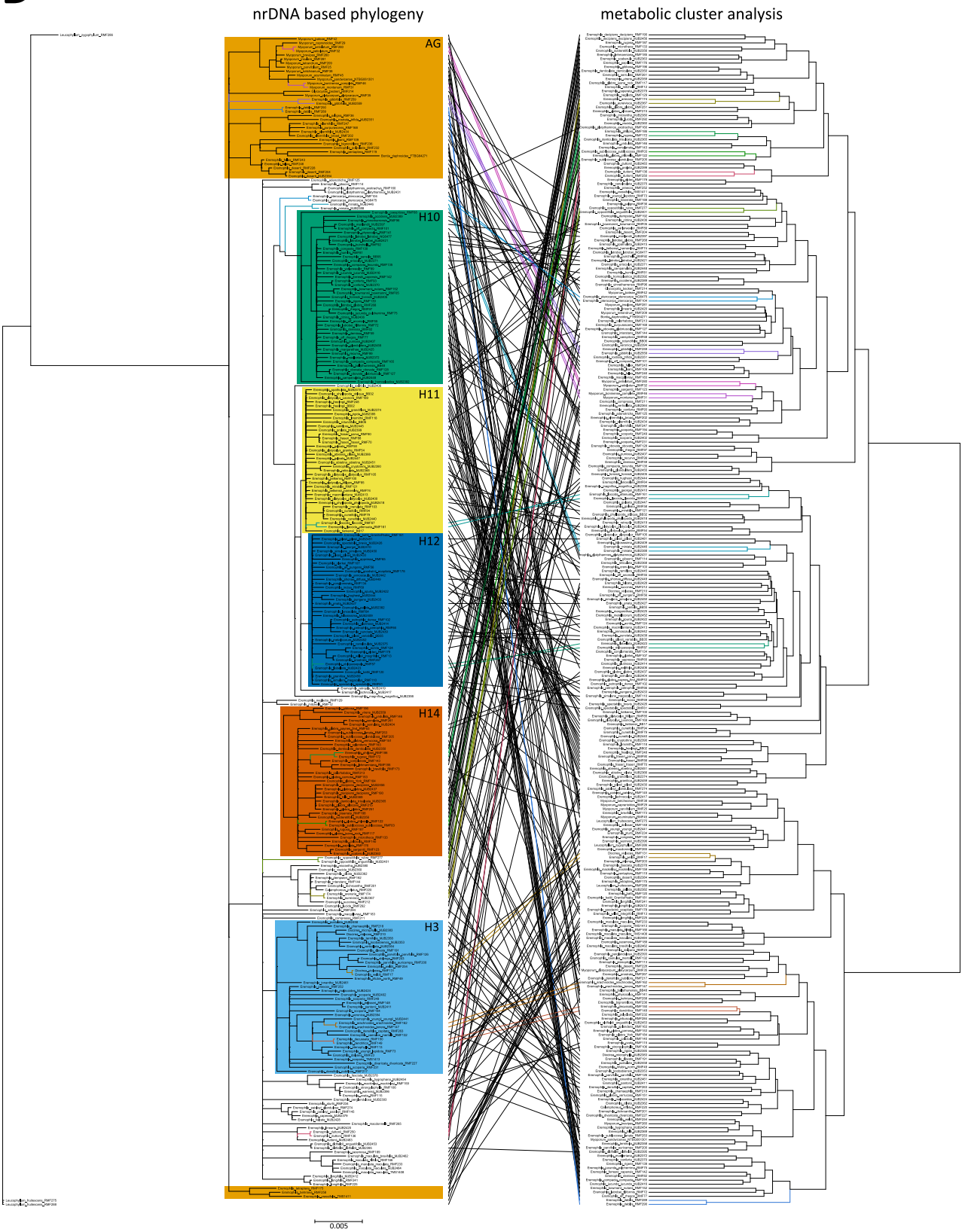


1095

C Tanglegram based on normalised 10696 spectral features (peak signal intensity)



D

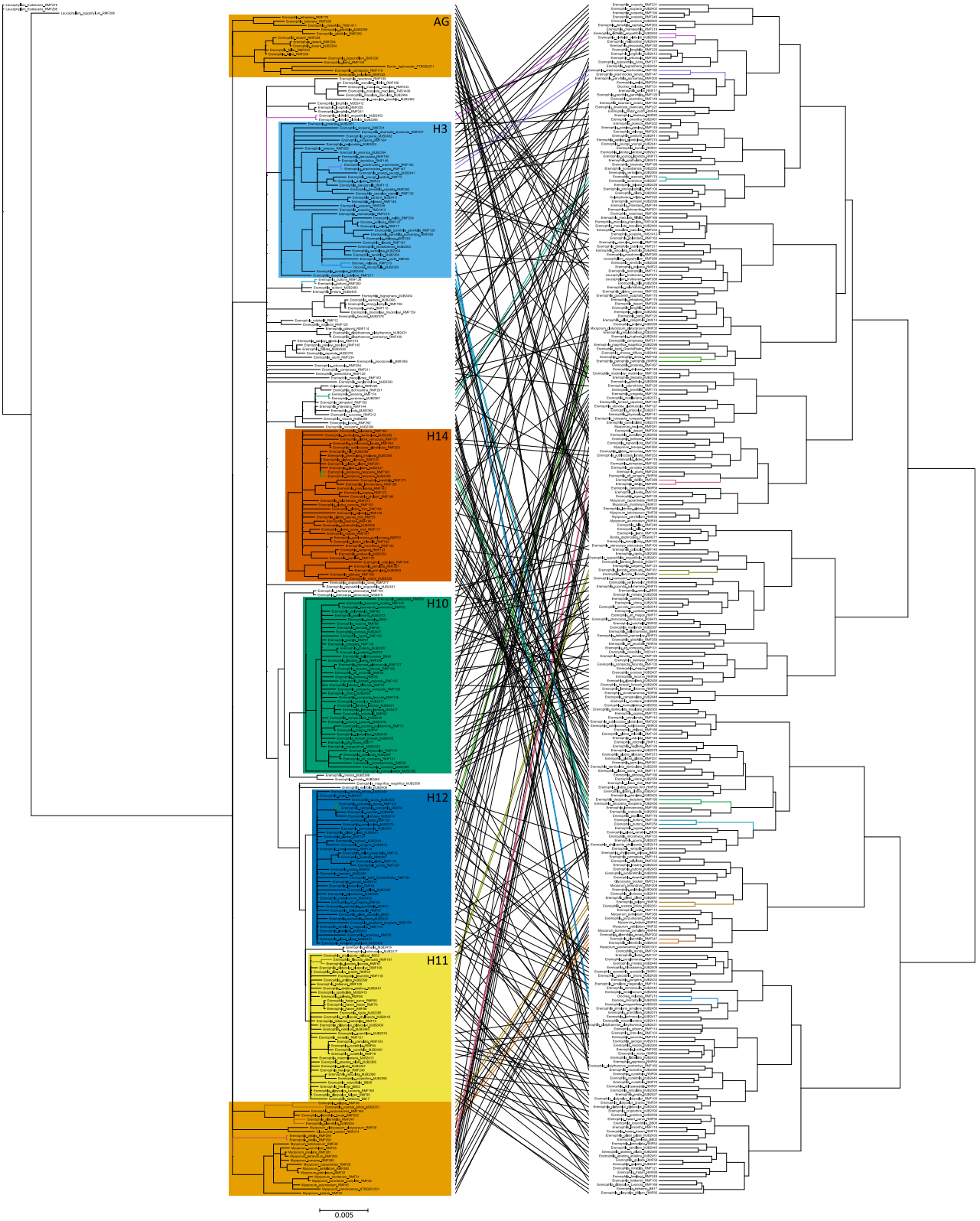

Tanglegram based on individual singleton spectral features
nrDNA based phylogeny
metabolic cluster analysis
AG
H10
H11
H12
H14
H3
0.005

E

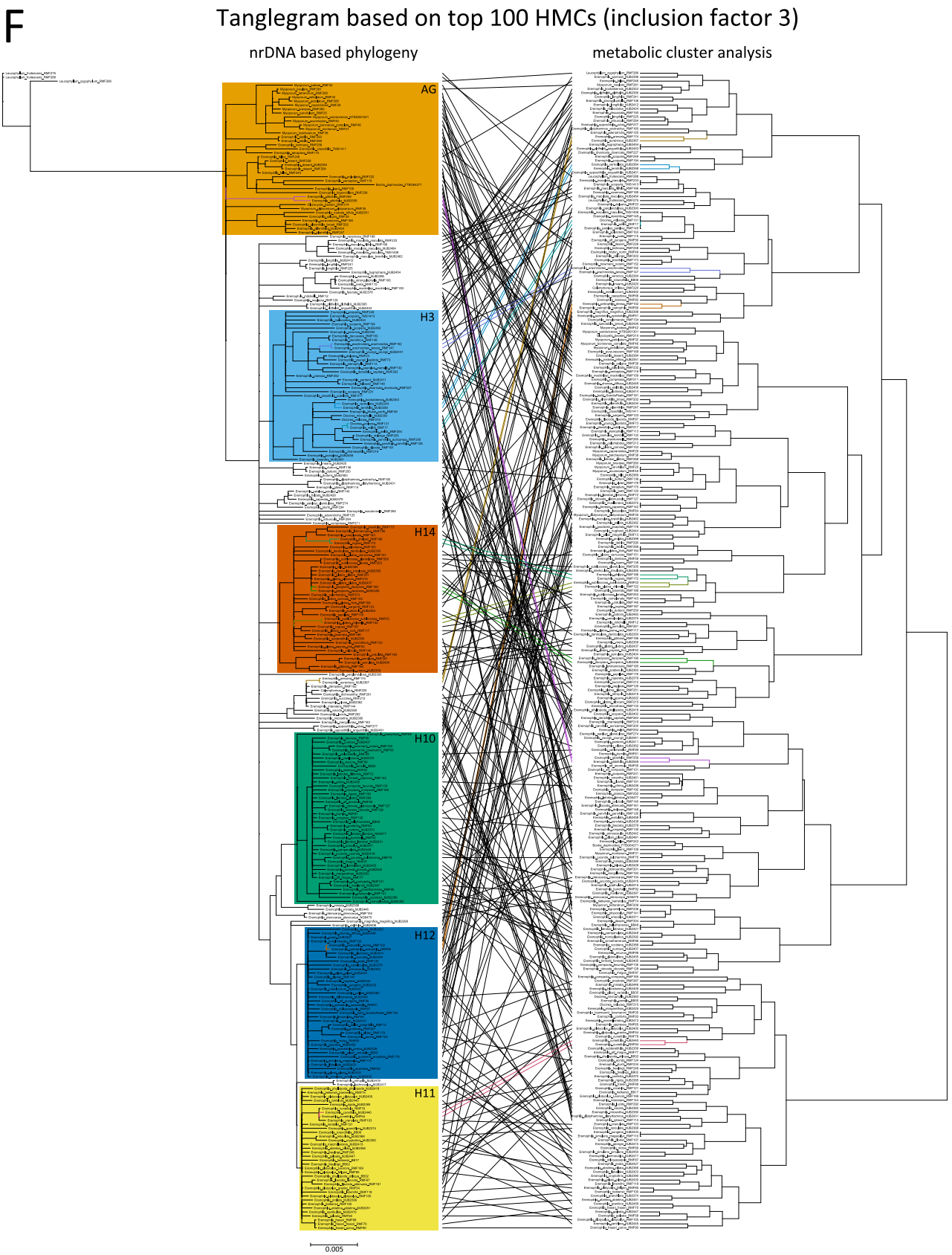
Tanglegram based on top 100 HMCs (inclusion factor 2)

nrDNA based phylogeny

metabolic cluster analysis



F


Tanglegram based on top 100 HMCs (inclusion factor 3)
nrDNA based phylogeny
metabolic cluster analysis
AG
H3
H14
H10
H12
H11
0.005

G

Tanglegram based on top 100 HMCs (inclusion factor 5)

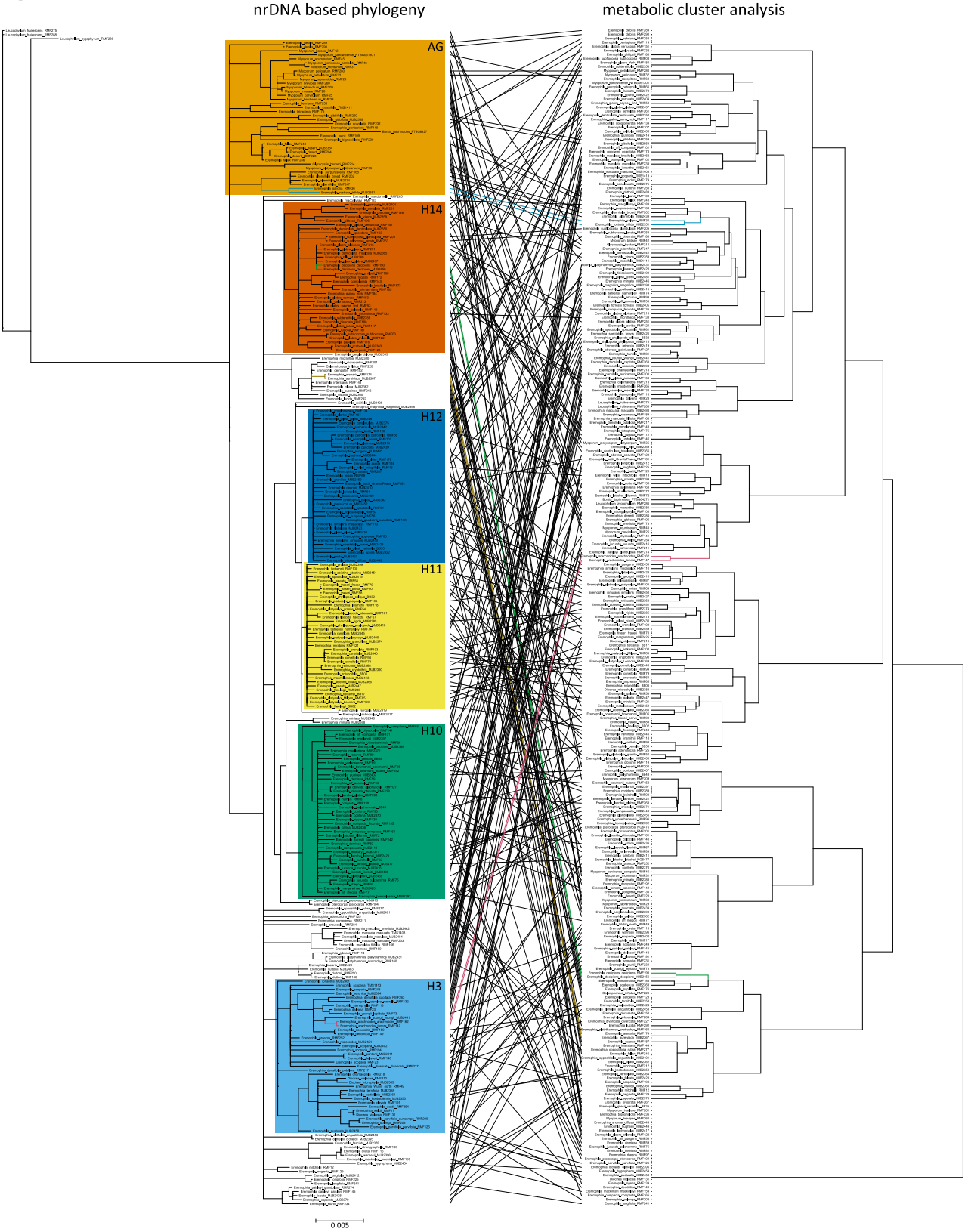


Figure S3. Collection of tanglegram analyses. Conjunction of phylogenetic and metabolic information from 291 species of the tribe Myoporeae visualized by different tanglegram analyses. The nrDNA based phylogeny shows taxa with posterior probability above 95%. Metabolic cluster analyses were conducted based on different datasets: (A) presence/absence of individual spectral features after normalisation of the full spectral dataset (10696 features). (B) presence/absence of the 30 chemical families found to be associated with serrulatane or viscidane diterpenoid metabolism within the generated global molecular network of Myoporeae. (C) peak signal intensity of individual spectral features after normalisation of the full spectral dataset (10696 features). (D) presence/absence of individual spectral features assigned as singletons after normalisation. (E) presence/absence of the 100 largest chemical families within the generated molecular network of Myoporeae using an inclusion factor of 2, (F) 3 and (G) 5. Phylogenetic subclades are indicated with the most prevalent ones highlighted by color code. The tanglegram analysis connects same specimens by a line, which is colored when equal branching of taxa is present in both analyses. Selected species with varying intraspecific chemical variation are highlighted within selected metabolic cluster analysis, with square (*E. duttonii*), diamond (*E. alternifolia*) and circle (*E. deserti*).

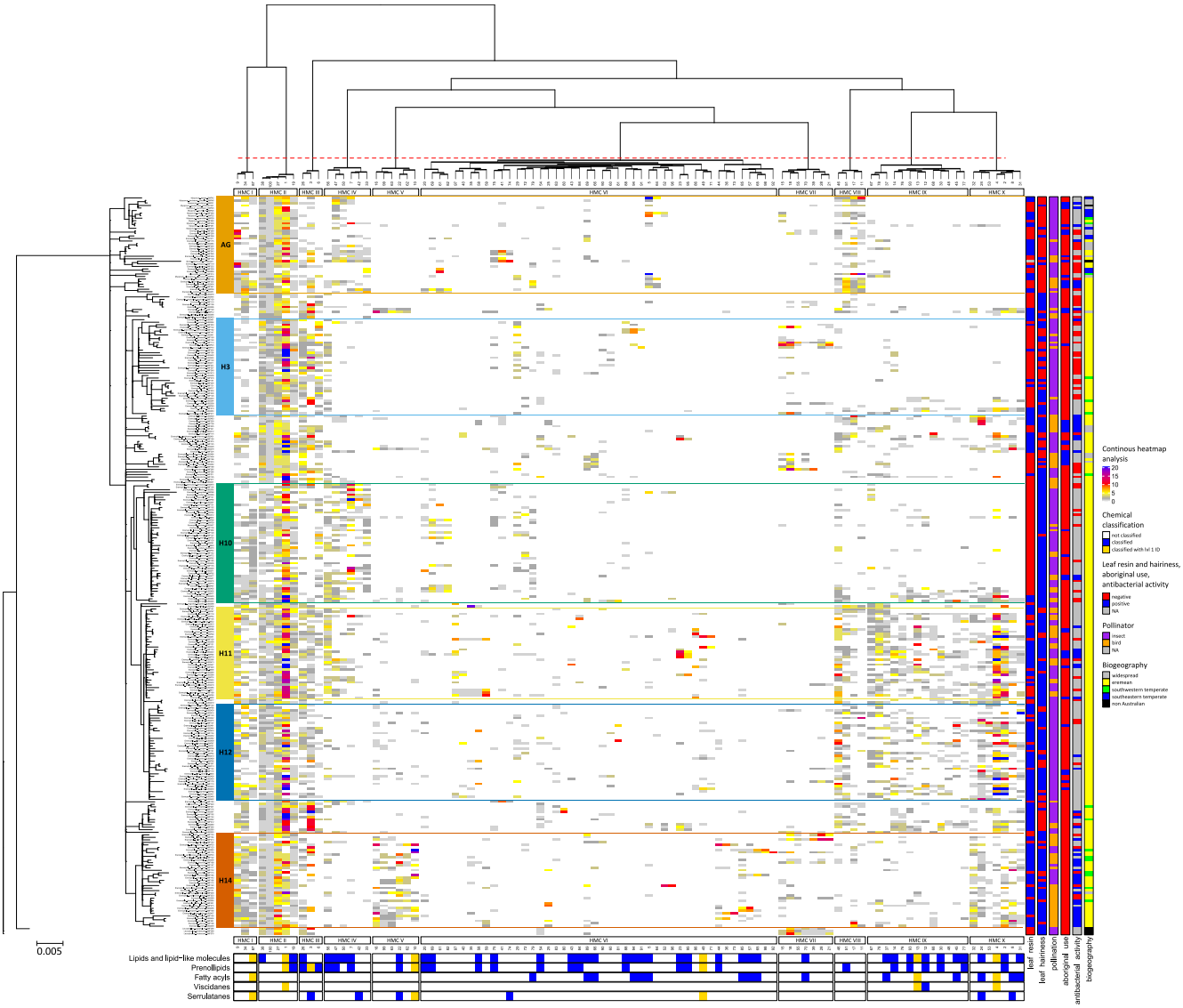


Figure S4. Continuous heatmap analysis showing chemo-evolutionary relationships in Myoporeae. A heatmap analysis was used to put chemical information into an evolutionary context. For that, information about the 100 largest chemical families was compiled, which derived from the molecular network of Myoporeae. For each specimen in this study, the total amount of metabolites of a corresponding chemical family is displayed within the heatmap. *LEFT*: The nuclear ribosomal DNA phylogeny on the left side presents the major phylogenetic clades. *TOP*: As depicted on top of the heatmap, the chemical information underwent a hierarchical clustering according to the given phylogeny to reveal chemo-evolutionary patterns among subsets of chemical families, defined as heatmap metabolic clusters (HMC) I – X. *BOTTOM*: Selected chemical classification generated by NAP based dereplication are displayed below in blue, while the presence of level 1 identification (m/z, retention time and MS2 match) in a particular chemical family is shown in gold. *RIGHT*: Functional annotations including the presence of leaf resin and hairiness, pollination, antibacterial activity, traditional medicinal usage as well as biogeographical species distribution information are displayed on the right side. A summary of legends are also included to the right of the figure.

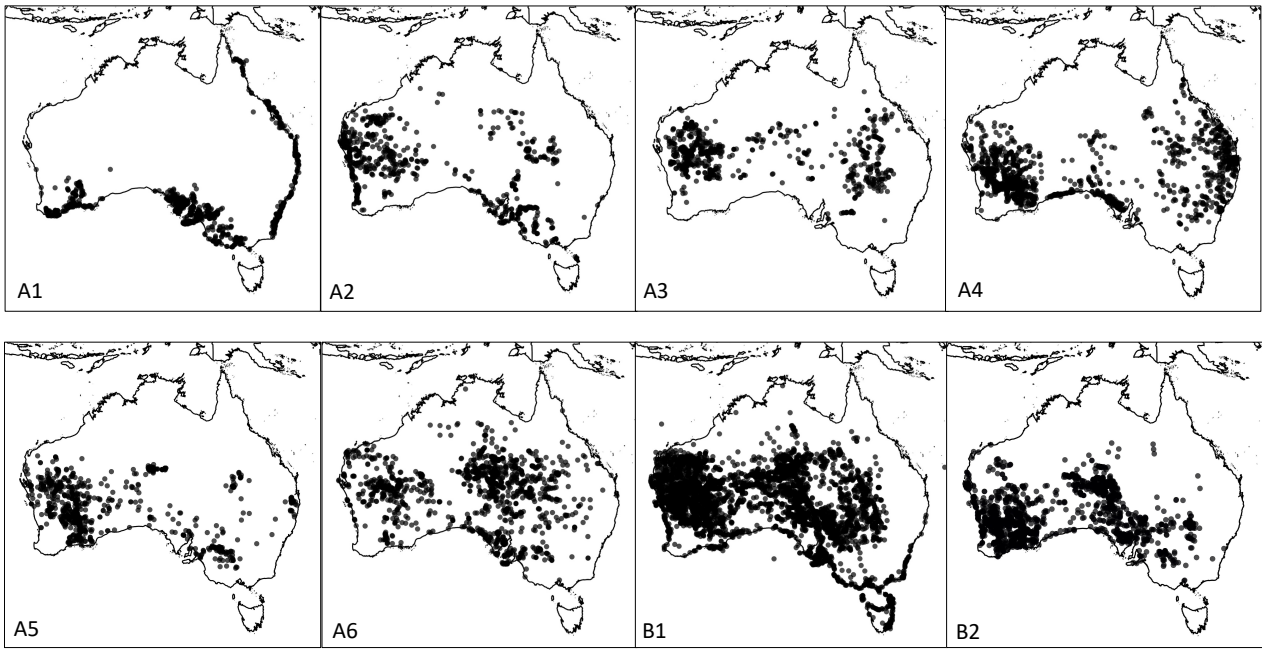


Figure S5. Comparison of species distribution between members of TMC A and B related chemistry. Species distributions for members of tanglegram metabolite clusters A (A1 – A6) and B (B1 – B2). Widespread species (those with broad distributions spanning more than one biome) have been excluded from maps. Species assessed as widespread were identified in clusters A1 (*Eremophila bignoniiflora*, *E. alternifolia*, *E. deserti*), A2 (*Myoporum montanum*, *E. latrobei* subsp. *glabra*, *E. deserti*), A5 (*M. acuminatum*, *E. longifolia*) and B2 (*E. mitchellii*). Species occurrence data was generated from the Australasian Virtual Herbarium (https://avh.chah.org.au/) as at 29/06/20. .

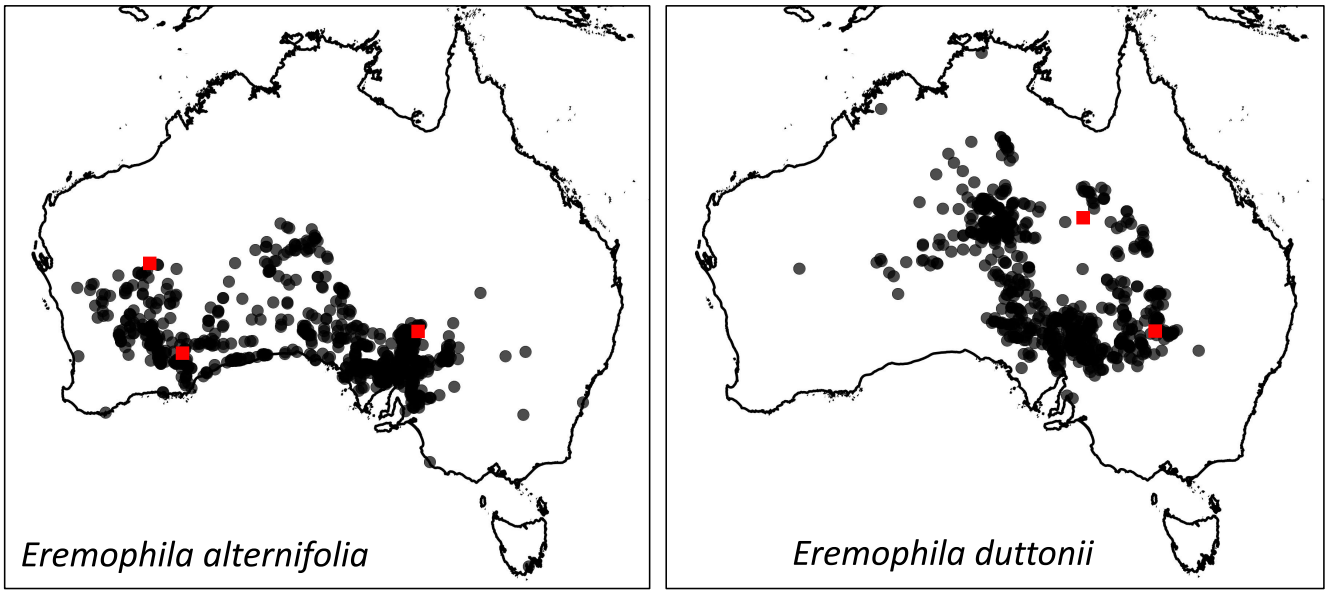


Figure S6. Geographic distributions of two widespread *Eremophila* species. *Eremophila alternifolia* and *Eremophila duttonii* specimen locations marked with red boxes. Data generated from the Australasian Virtual Herbarium (https://avh.chah.org.au/) as at 06/04/20.

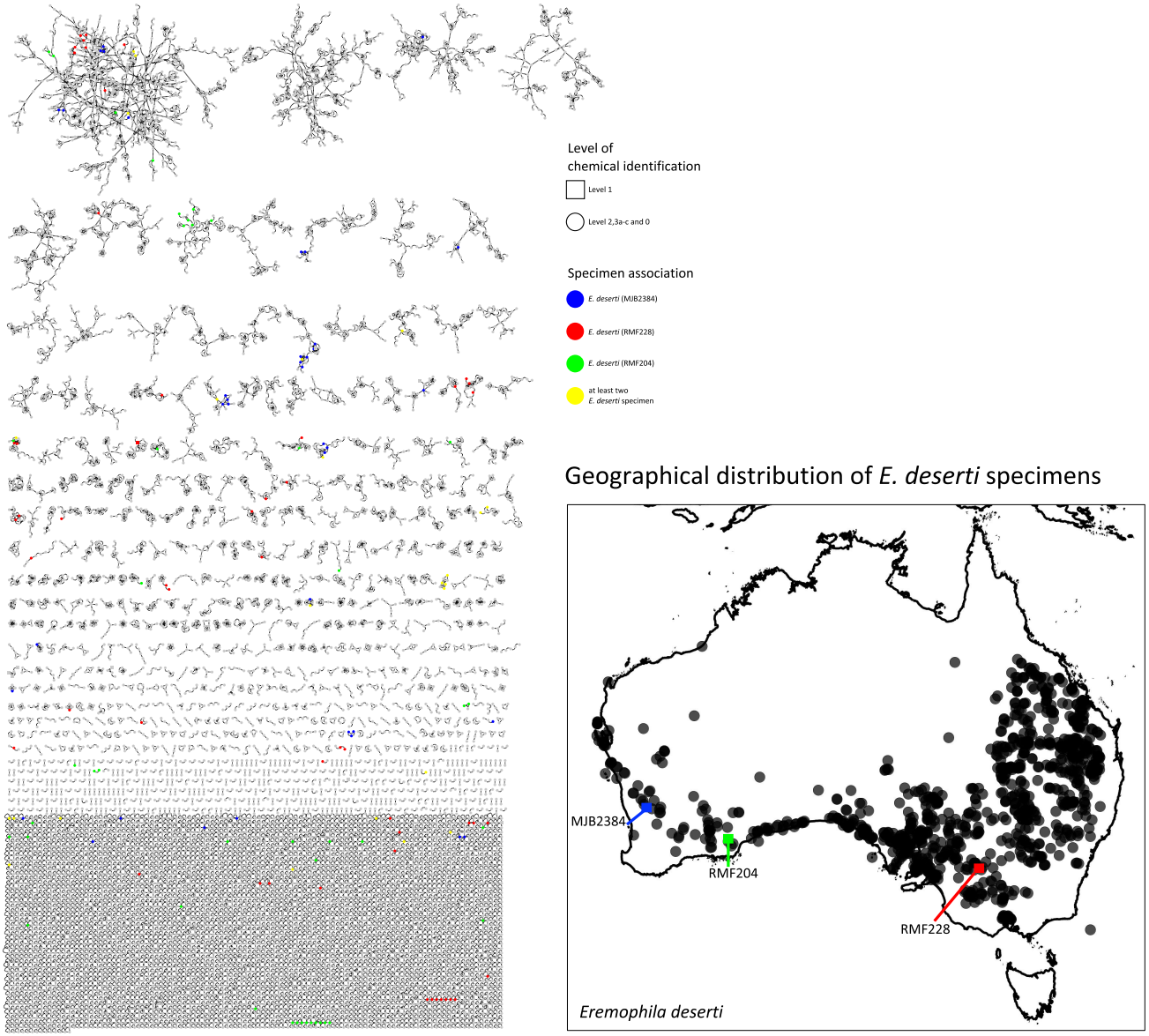


Figure S7. Molecular network analysis of *Eremophila deserti*. Global molecular network representation of metabolite distribution from three *E. deserti* specimens. Highlighted in yellow are putative metabolites that are shared by at least two specimen of *E. deserti*. Metabolites found specifically in one of the specimen are marked in blue: *E. deserti* (MJB2384), red: *E. deserti* (RMF228) and green: *E. deserti* (RMF204). On each node in the network, the total number of specimen that share this specific metabolite is displayed. Indicated on the right side is the geographic distribution of the observed *E. deserti* specimen in this study, with corresponding sampling locations marked with colored boxes. Species occurrence data was generated from the Australasian Virtual Herbarium (https://avh.chah.org.au/) as at 06/04/20. Maps of Australian annual mean temperature, annual mean precipitation and annual mean solar radiation were downloaded as layers for reference from the Atlas of Living Australia (ala.org.au)

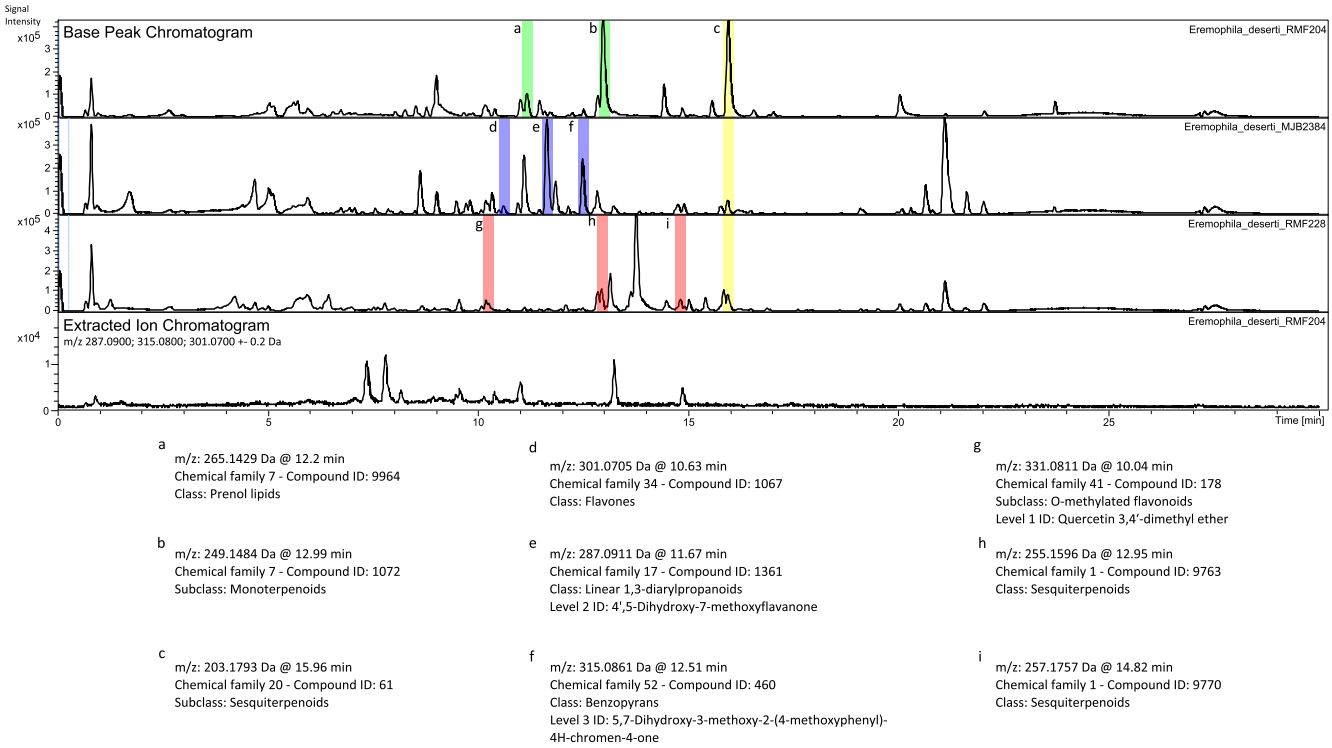


Figure S8. Spectral comparison of three *Eremophila deserti* specimens. Spectral representation of base peak chromatograms for three *Eremophila deserti* specimens that display interspecific chemical variation. Selected compounds are highlighted and chemical annotation stated below. An extracted ion chromatogram of specimen RMF204 is also shown to show the complete lack of flavonoid compounds present in specimen MJB2384.

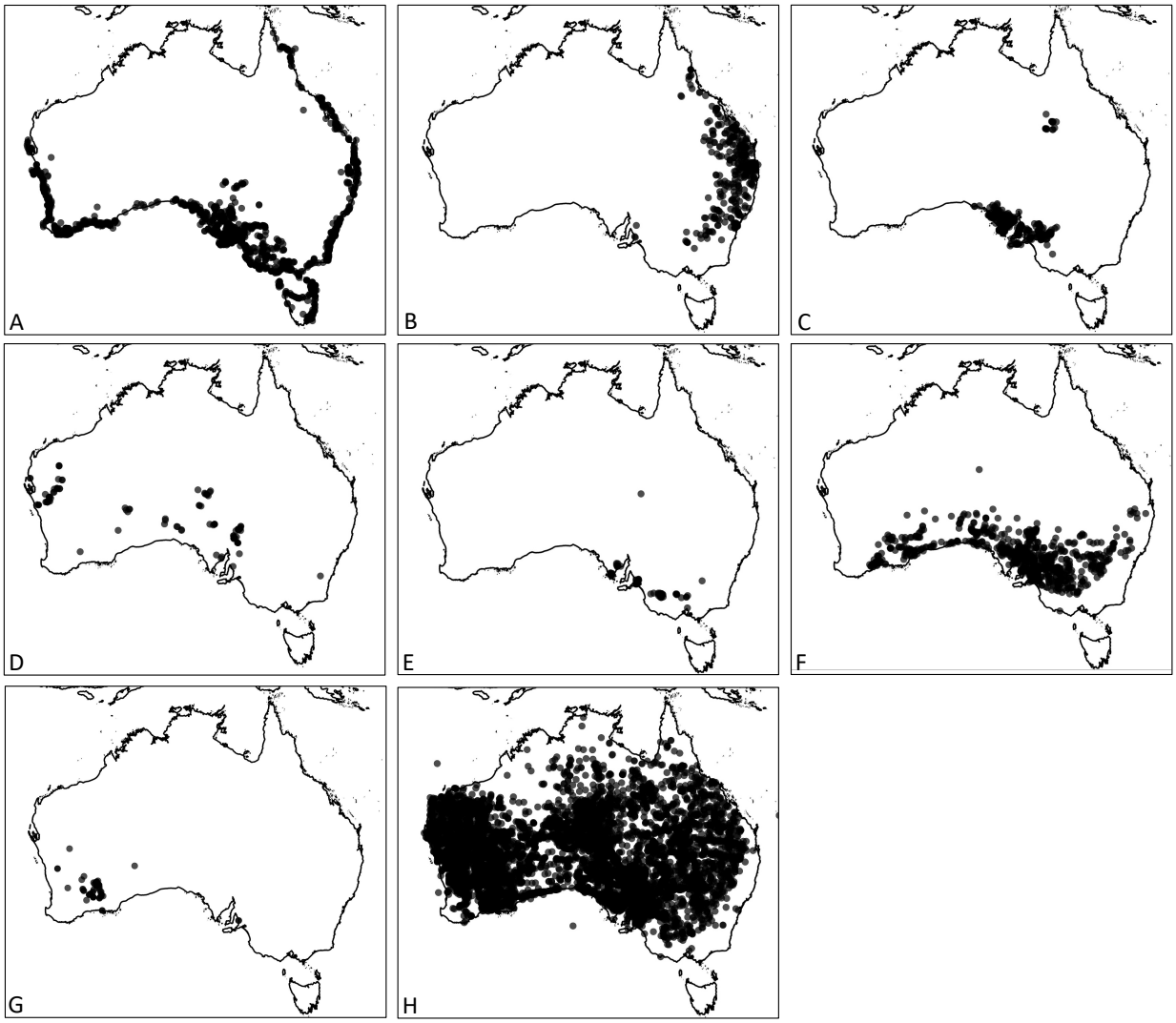


Figure S9. Species distributions for members of Myoporeae phylogenetic clades A–H. In this representation, widespread species (those with broad distributions spanning more than one biome) have been excluded from maps. Species assessed as widespread were identified in phylogenetic clades A (*M. acuminatum*, *M. montanum*) D (*E. bignoniiflora*, *E. deserti*, *E. polyclada*), G (*E. alternifolia*) and H (*E. latrobei* subsp. *glabra*, *E. longifolia*, *E. maculata* subsp. *maculata*, *E. mitchellii*). Species occurrence data was generated from the Australasian Virtual Herbarium (https://avh.chah.org.au/) as at 06/04/20.

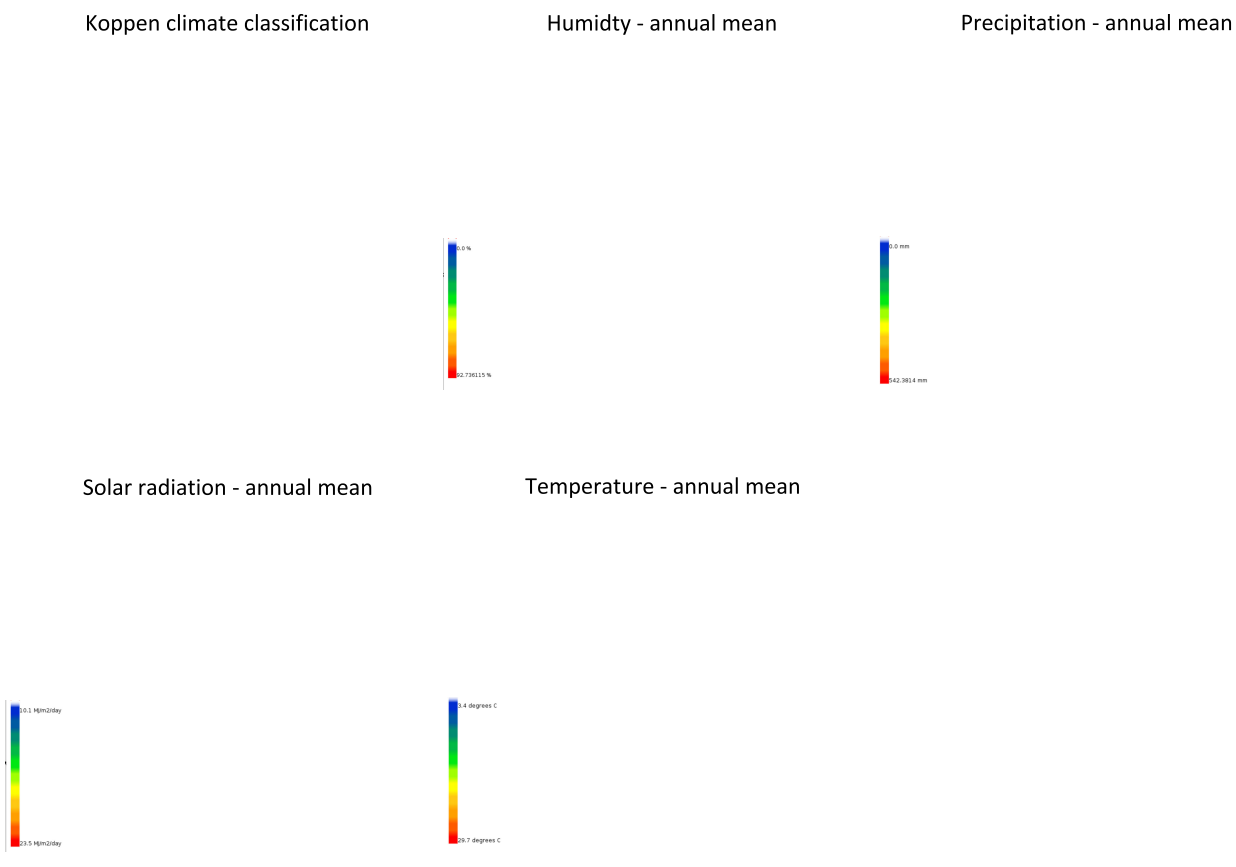


Figure S10. Climatic information about Australia. Maps of Australian annual mean temperature, annual mean precipitation, annual mean solar radiation, annual mean relative humidity as well as major classes of the Koppen Climate Classification were downloaded as layers for reference towards the species distribution maps in this study from the Atlas of Living Australia (ala.org.au).

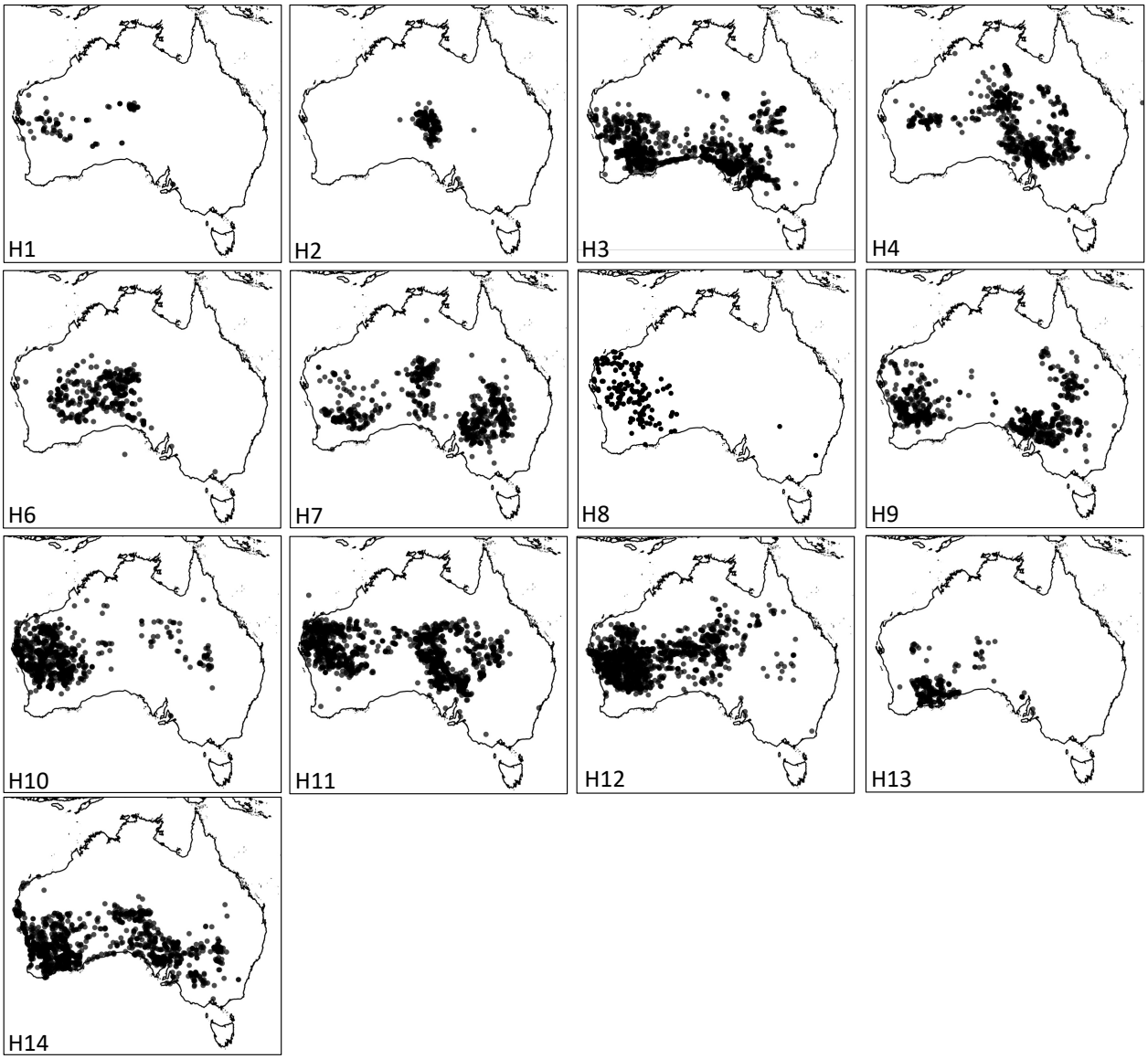


Figure S11. Species distributions for members of Myoporeae phylogenetic clade H, subclades H1–H14. In this representation, widespread species (those with broad distributions spanning more than one biome) have been excluded from maps. Species assessed as widespread were identified in phylogenetic subclades H2 (*E. mitchellii*), H5 (contains only *E. longifolia*, so no distribution map included), H8 (*E. maculata* subsp. *maculata*) and H10 (*E. latrobei* subsp. *glabra*). Species occurrence data was generated from the Australasian Virtual Herbarium (https://avh.chah.org.au/) as at 06/04/20.

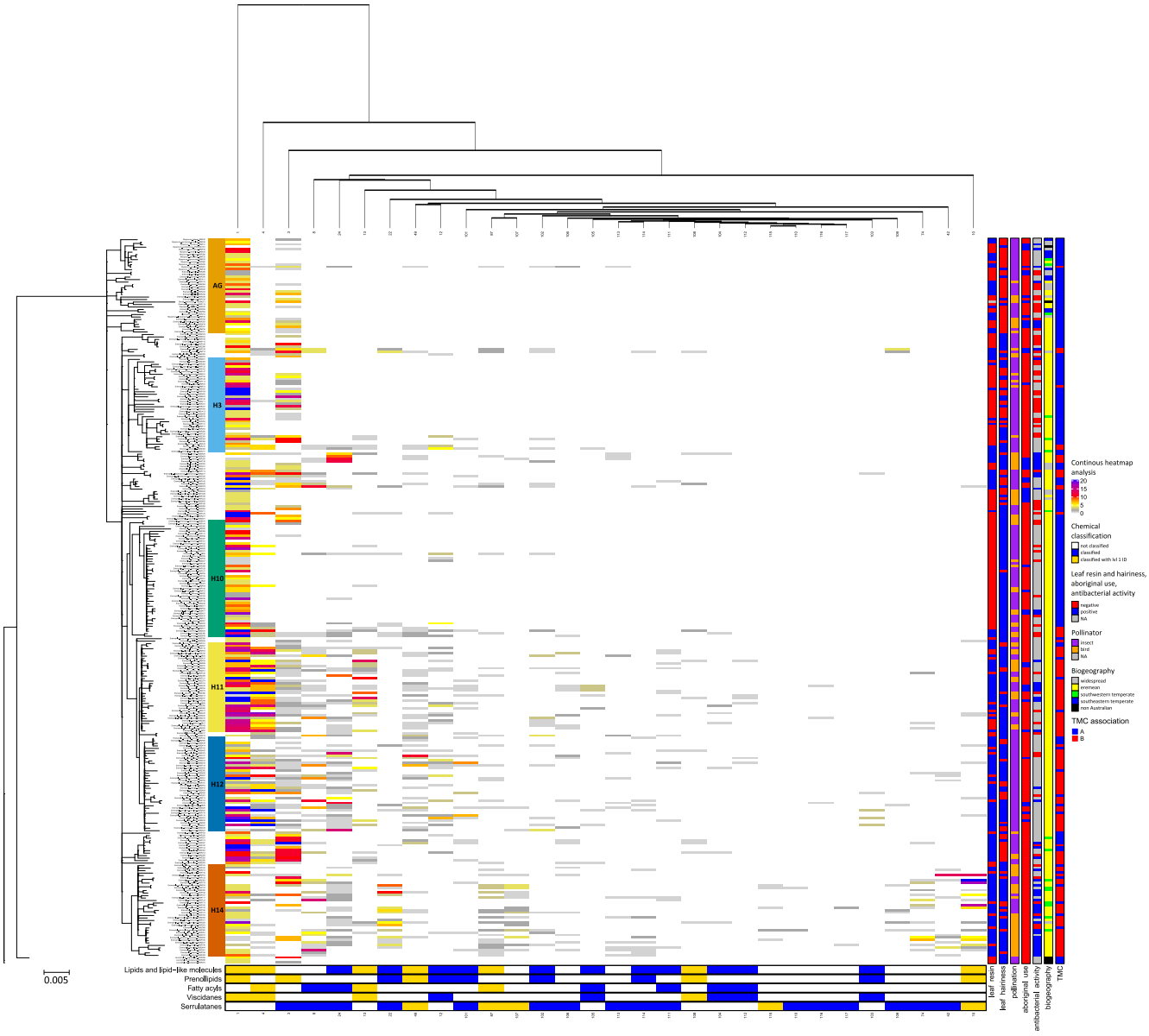


Figure S12. Continuous heatmap analysis of chemo-evolutionary patterns involving serrulatane and viscidane diterpenoids in Myoporeae. A heatmap analysis was used to put chemical information into an evolutionary context. For that, information about the 30 chemical families derived from the molecular network of Myoporeae, which have been found to be associated with serrulatane and viscidane diterpene chemistry was compiled. For each specimen in this study, the total amount of metabolites of a corresponding chemical family is displayed within the heatmap. *LEFT*: The nuclear ribosomal DNA phylogeny on the left side presents the major phylogenetic clades. *TOP*: As depicted on top of the heatmap, the chemical information underwent a hierarchical clustering according to the given phylogeny to reveal chemo-evolutionary patterns. *BOTTOM*: Selected chemical classification generated by NAP based dereplication are displayed below in blue, while the presence of level 1 identification (m/z, retention time and MS^2 match) in a particular chemical family is shown in gold. *RIGHT*: Functional annotations including the presence of leaf resin and hairiness, pollination, antibacterial activity, traditional medicinal usage, biogeographical species distribution information as well as tanglegram metabolic cluster (TMC) association are displayed on the right side. A summary of legends are also included to the right of the figure.

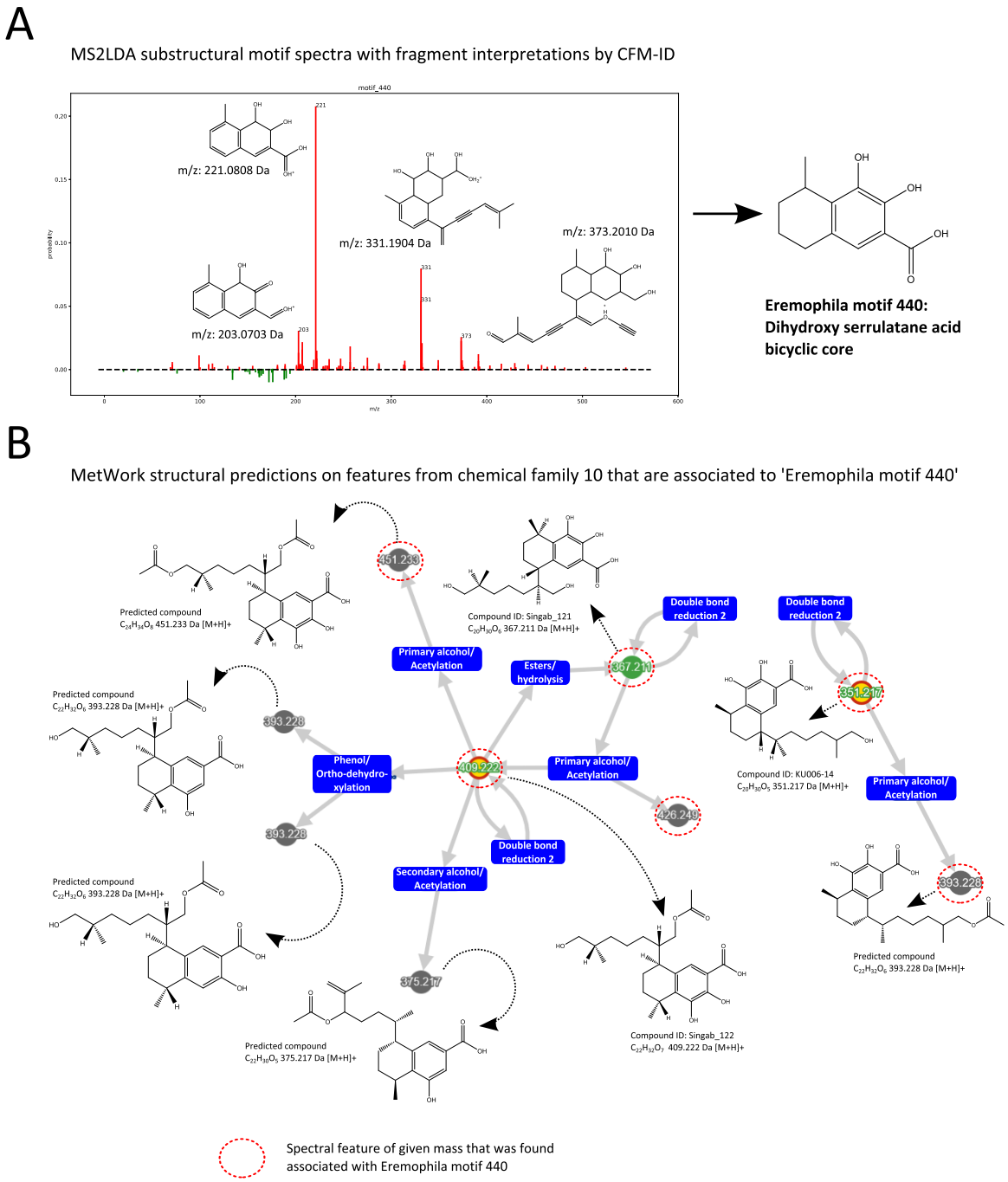


Figure S13. Annotation of substructural motif 'Eremophila motif 440'. Annotation of the unknown substructural motif 'Eremophila motif 440' was conducted using the combined information from the biochemical reaction prediction tool MetWork and the *in silico* MS^2 spectra prediction tool CFM-ID. (A) The spectral representation of the studied MS2LDA motif indicates the probability of presence and absence of distinct fragments, which have been annotated using *in silico* spectra prediction by CFM-ID on serrulatane diterpenoid structures found to be associated with 'Eremophila motif 440', such as KU006-14. Based on these structural annotations a dihydroxy serrulatane acid bicyclic core was proposed as the substructure underlying this motif. (B) MetWork based prediction of biochemical and structural reactions joining serrulatane diterpenoid features present in chemical family 10 that are associated with 'Eremophila motif 440'. The predictions indicate the loss of the motif upon dehydroxylation at the aromatic ring, underlying these specific modifications as the inherent motif characteristics. Nodes highlighted in light green/yellow correspond to prior dereplicated spectral features, while dark green displays unknown features that where structurally annotated by the MetWork analysis.

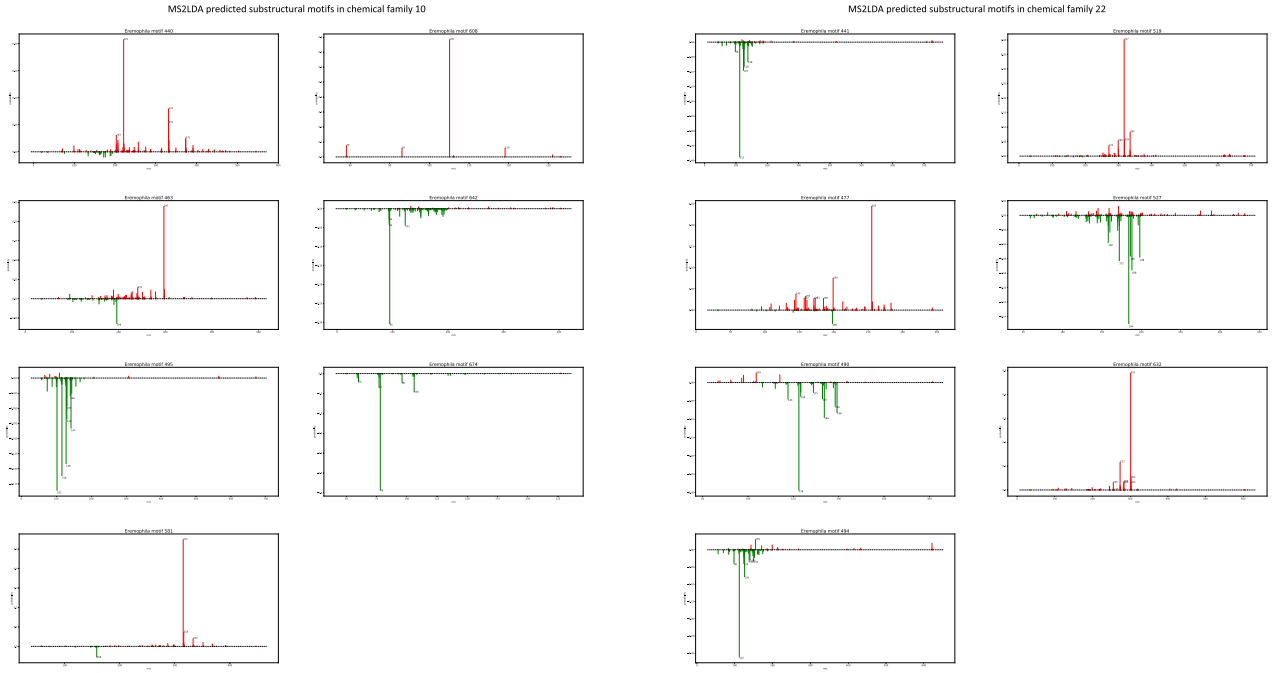


Figure S14. Spectral representation of MS2LDA predicted substructural motifs found in chemical family 10 and 22. Spectral information on the probability of mass fragments present or absent within different substructural motifs determined in chemical families 10 and 22 by the MS2LDA tool.

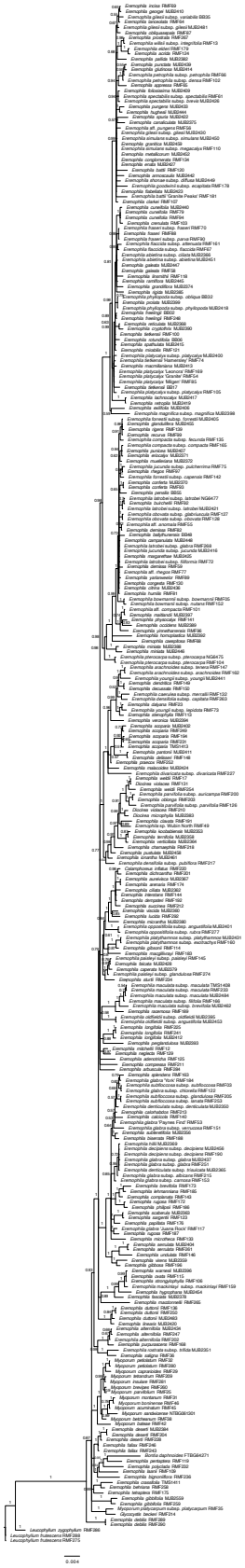


Figure S15. Phylogeny of tribe Myoporeae based on analysis of nuclear ribosomal DNA sequences. Bayesian inference 50% majority-rule consensus tree, showing posterior probability values on each branch. Branches with values <0.50 collapsed.

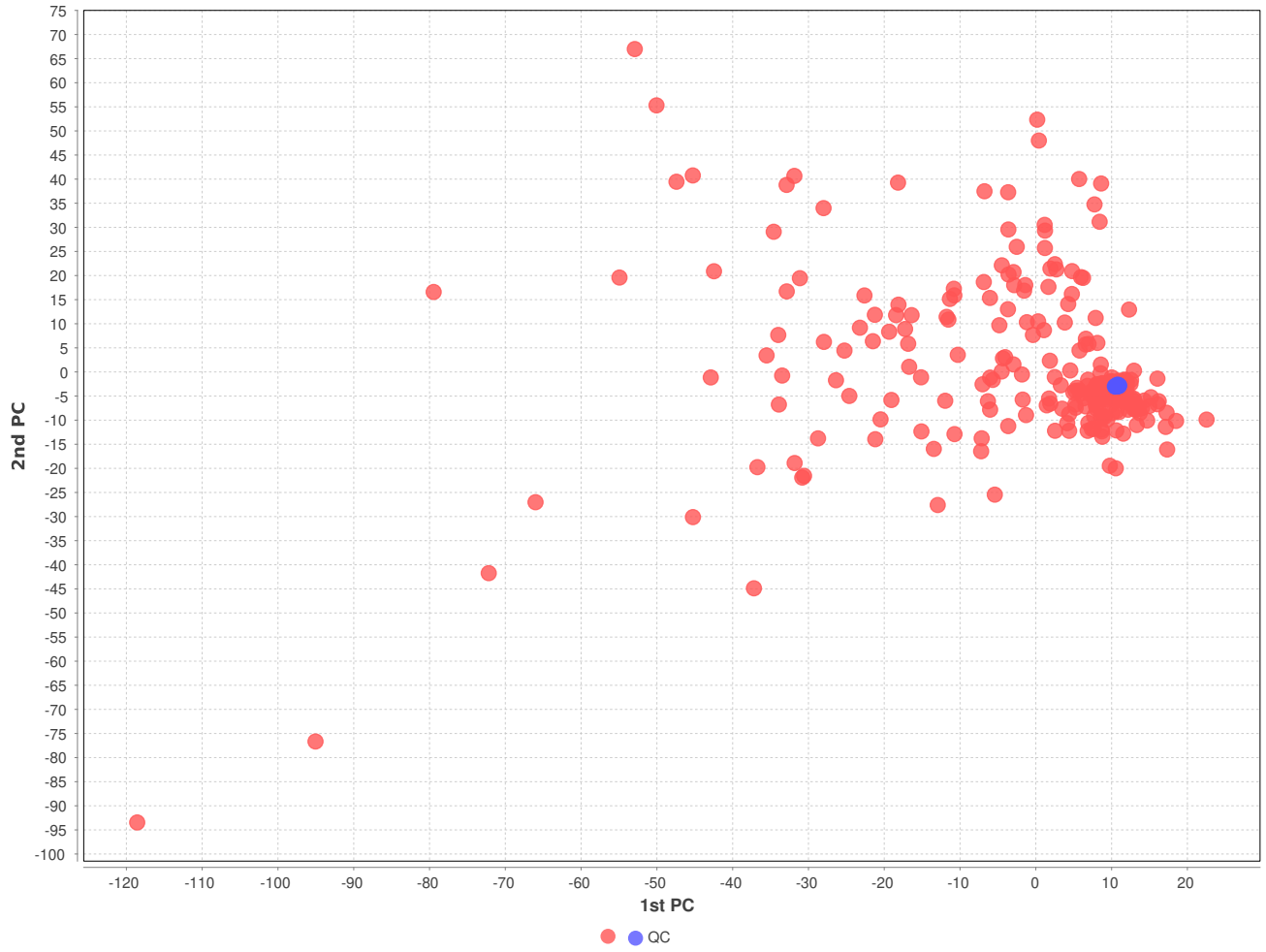


Figure S16. LC-qToF-HRMS data quality control. Assessing quality of the full spectral dataset used in this study by plotting all samples (red) together with the quality controls (blue) within a principal component analysis.