New Phytologist Supporting Information

Article title: Quantifying chemodiversity considering biochemical and structural properties of compounds with the R package *chemodiv*
Authors: Hampus Petrén, Tobias G. Köllner, Robert R. Junker
Article acceptance date: 11 December 2022

The following Supporting Information is available for this article:

Dataset S1 (see separate file) Dataset used for all analyses and figures in the manuscript.

Notes S1 (see separate file) R-script used to perform statistical analyses and create figures in the manuscript.

Fig. S1 Comparisons of different measures of phytochemical diversity and dissimilarity for the semi-simulated dataset with cardenolides and glucosinolates.

Fig. S2 Comparison of glucosinolate diversity and dissimilarity, calculated with and without taking compound dissimilarities into account, for 48 *Erysimum* species.

Fig. S3 Comparisons of different measures of phytochemical diversity for eight groups of phytochemical samples simulated to have a high or low richness, evenness and compound dissimilarity.

Fig. S4 Comparisons of compound dissimilarities calculated using the three different methods in the *compDis* function.

Fig. S5 Dendrogram showing compound dissimilarities for floral scent compounds found in *Achillea millefolium* and *Cirsium arvense*.

Fig. S6 Computation times of the *quickChemoDiv* function for datasets with varying number of compounds, computed with and without calculation of compound dissimilarity data.

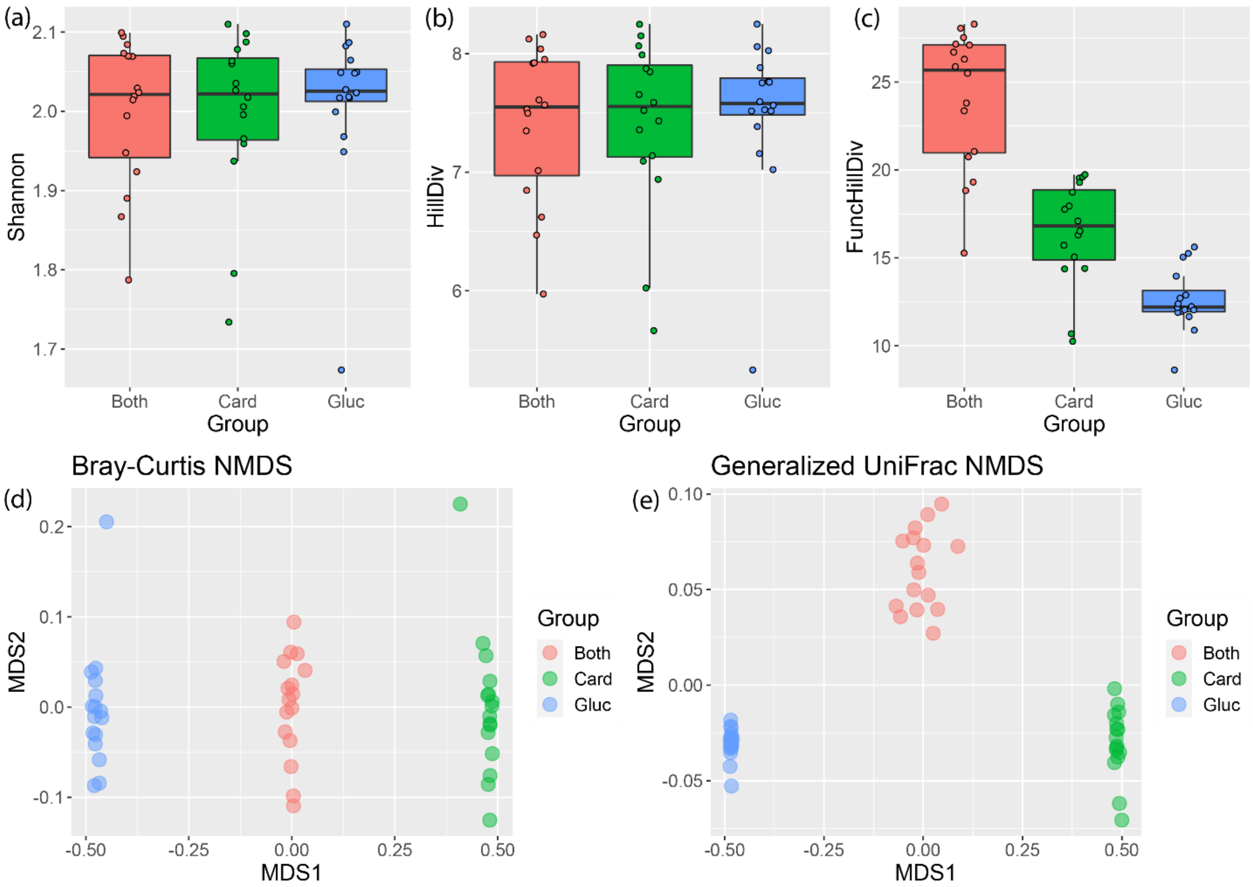

Fig. S1 Comparison of different measures of phytochemical diversity (a-c) and dissimilarity (d-e) for the semi-simulated dataset with cardenolides and glucosinolates. For measures of phytochemical diversity, Shannon's diversity (a) and Hill diversity (b) (equal to the exponential of Shannon's diversity) do not take compound dissimilarity into account, and all three groups have similar diversity. In contrast, functional Hill diversity (c) depends also on compound dissimilarity, with the result that the group with a high concentration of both cardenolides and glucosinolates (red) has a higher phytochemical diversity than the groups with only cardenolides (green) or only glucosinolates (blue) at high concentrations. Boxes display median values and upper and lower quartiles, with whiskers extending up to 1.5 times the inter-quartile range. For measures of phytochemical dissimilarity, when Bray-Curtis dissimilarities are used, the within-group dispersion is similar in all three groups. In contrast, when Generalized UniFrac dissimilarities are used, which take compound dissimilarity into account, the group with both types of compounds at a high concentration has a higher within-group dispersion, as an effect of containing a mixture of more dissimilar compounds.

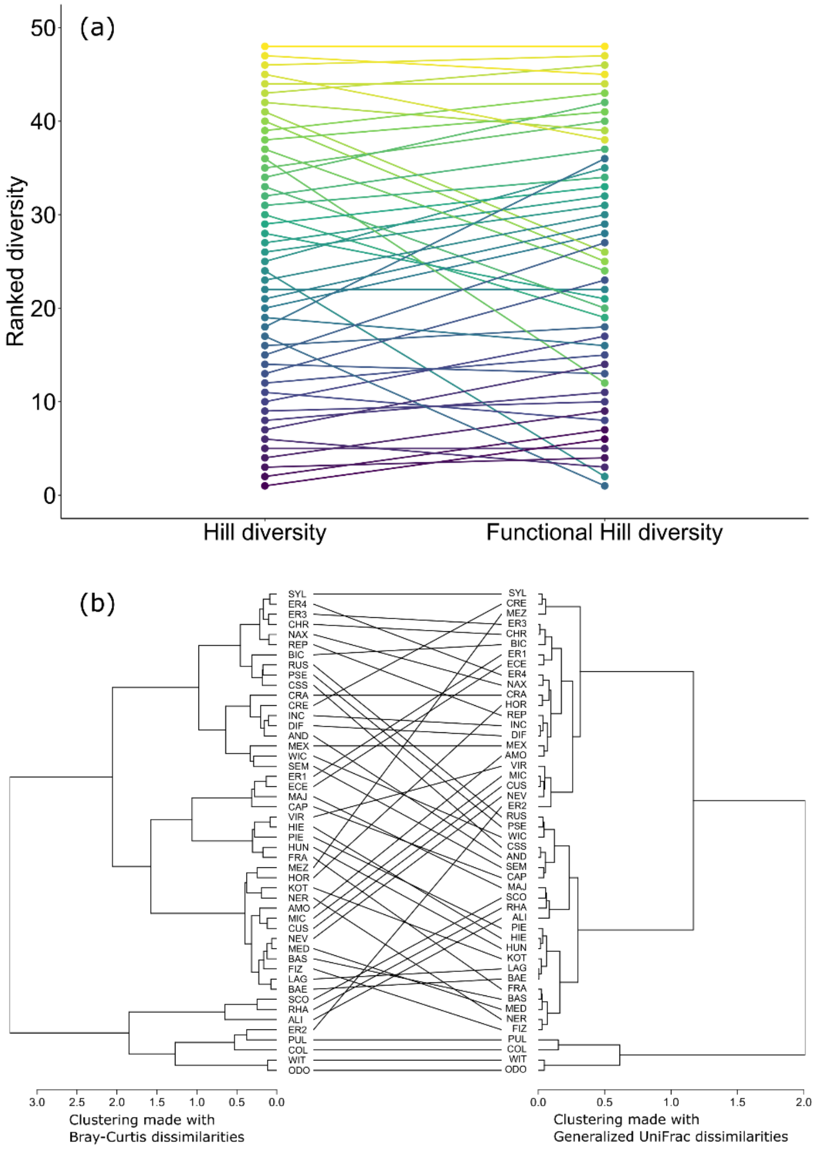

Fig. S2 Comparison of glucosinolate diversity and dissimilarity for 48 *Erysimum* species, calculated with and without taking compound dissimilarities into account. (a) Glucosinolate diversity measured as Hill diversity (left), and functional Hill diversity (right). (b) Glucosinolate dissimilarity for the same dataset, visualized with dendrograms constructed from Bray-Curtis dissimilarities (left) or Generalized UniFrac dissimilarities (right). Clustering was done with Ward's minimum variance method to match Züst *et al.* (2020). Functional Hill diversity and Generalized UniFracs take compound dissimilarities into account. In both (a) and (b), patterns partly differ for analyses including/excluding compound dissimilarities, with an overall moderate effect on results. For functional Hill diversity and Generalized UniFracs, compound dissimilarities were calculated based on *PubChem Fingerprints*. Data from Züst *et al.* (2020).

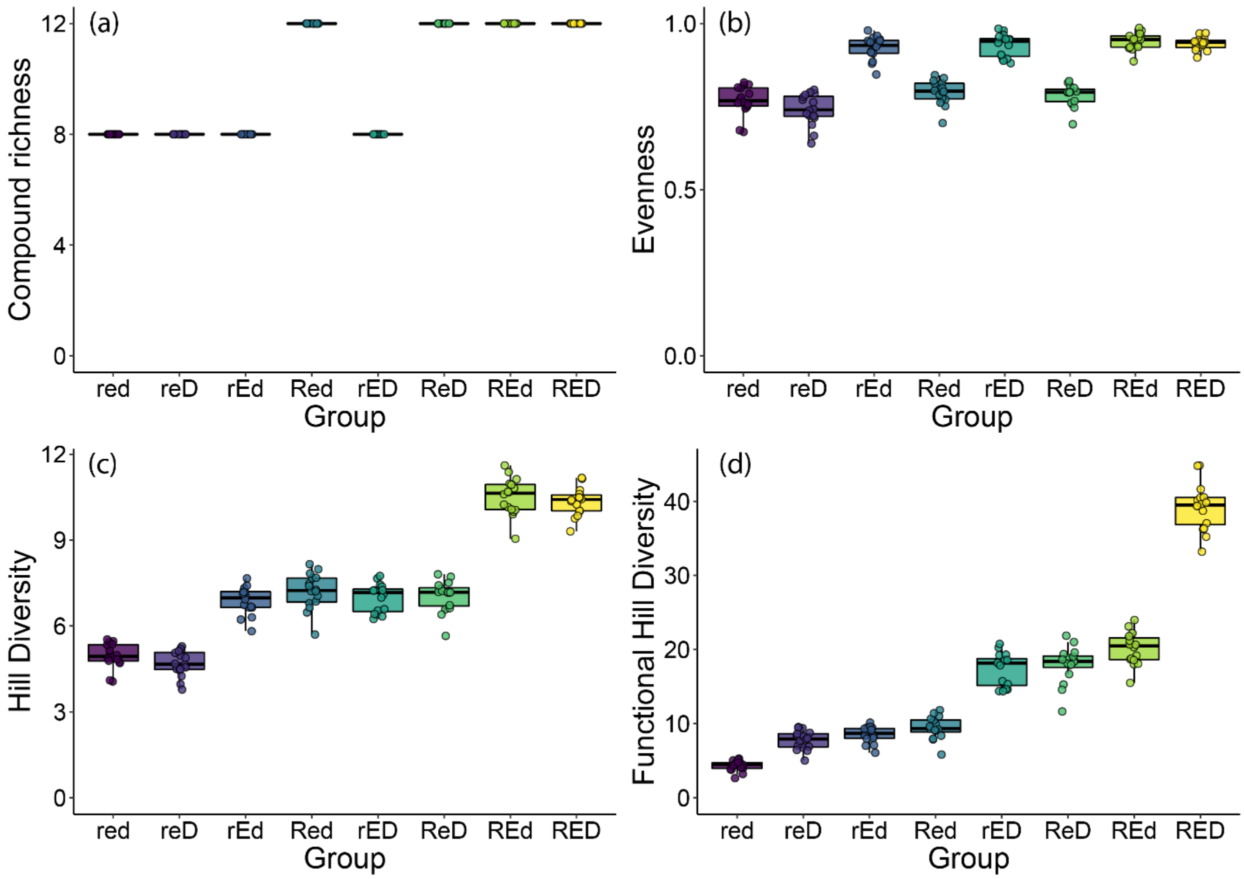

Fig. S3 Compound richness (a), Pielou's evenness (b), Hill diversity (c) and functional Hill diversity (d) for eight different groups of simulated phytochemical samples. Each group consists of 16 samples simulated to have a high or low richness (r/R), evenness (e/E) and compound dissimilarity (d/D). Lowercase letters indicate a low value, uppercase letters indicate a high value. Richness (a) is the number of compounds in the samples (and is equal to Hill diversity at *q* = 0). Evenness (b) depends on the relative abundances of compounds. Hill diversity (*q* = 1) (c), equal to the exponential of Shannon's diversity, depends on both richness and evenness, and is therefore higher for groups with high richness and/or evenness. Functional Hill diversity (*q* = 1) (d) is dependent on all three components of diversity (richness, evenness and disparity). It is lowest when richness, evenness and disparity is low, intermediate when one or two of the components is high, and highest when richness, evenness and dissimilarity are all high. In this regard, functional Hill diversity is therefore the most comprehensive measure of diversity. Boxes display median values and upper and lower quartiles, with whiskers extending up to 1.5 times the inter-quartile range.

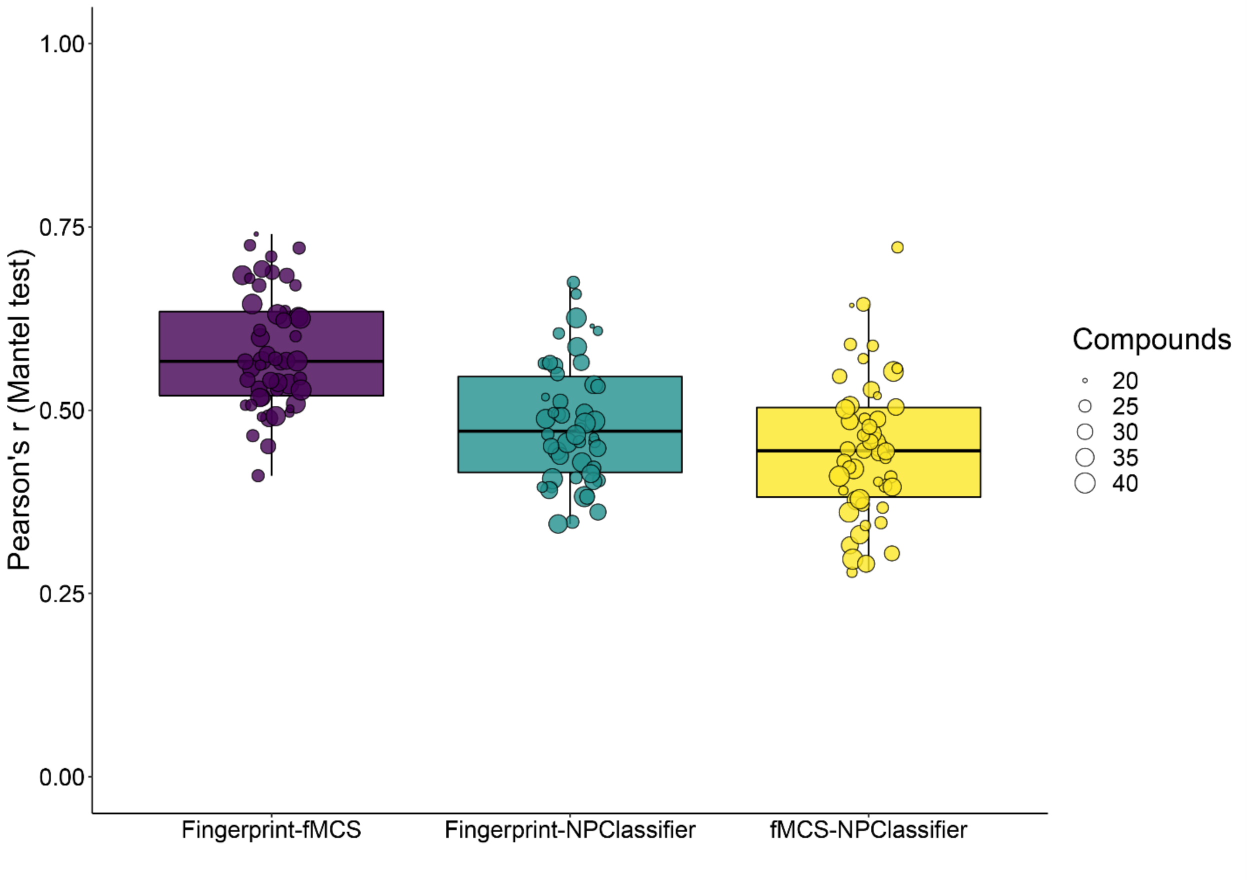

Fig. S4 Comparison of compound dissimilarities calculated using the three different methods in the *compDis* function. In 50 iterations, 20-40 phytochemical compounds were randomly selected, and compound dissimilarities were calculated using three methods: *NPClassifier*, which compares compounds based on a classification of compounds into groups largely corresponding to biosynthetic pathways, and *PubChem fingerprints* and *fMCS*, which compare compounds based on structural properties of the molecules using binary fingerprints and substructure matching, respectively. Mantel tests were then used to calculate Pearson's correlation coefficients between dissimilarity matrices. Boxes display median values and upper and lower quartiles, with whiskers extending up to 1.5 times the inter-quartile range. Data points of individual comparisons are overlaid. Correlations were overall moderately strong, and statistically significant ($P < 0.05$) in all cases. Mean correlation coefficients were highest between dissimilarity matrices based on *PubChem fingerprints* and *fMCS* (mean $r = 0.58$), and somewhat lower between dissimilarity matrices based on *PubChem fingerprints* and *NPClassifier* (mean $r = 0.48$) and dissimilarity matrices based on *fMCS* and *NPClassifier* (mean $r = 0.45$).

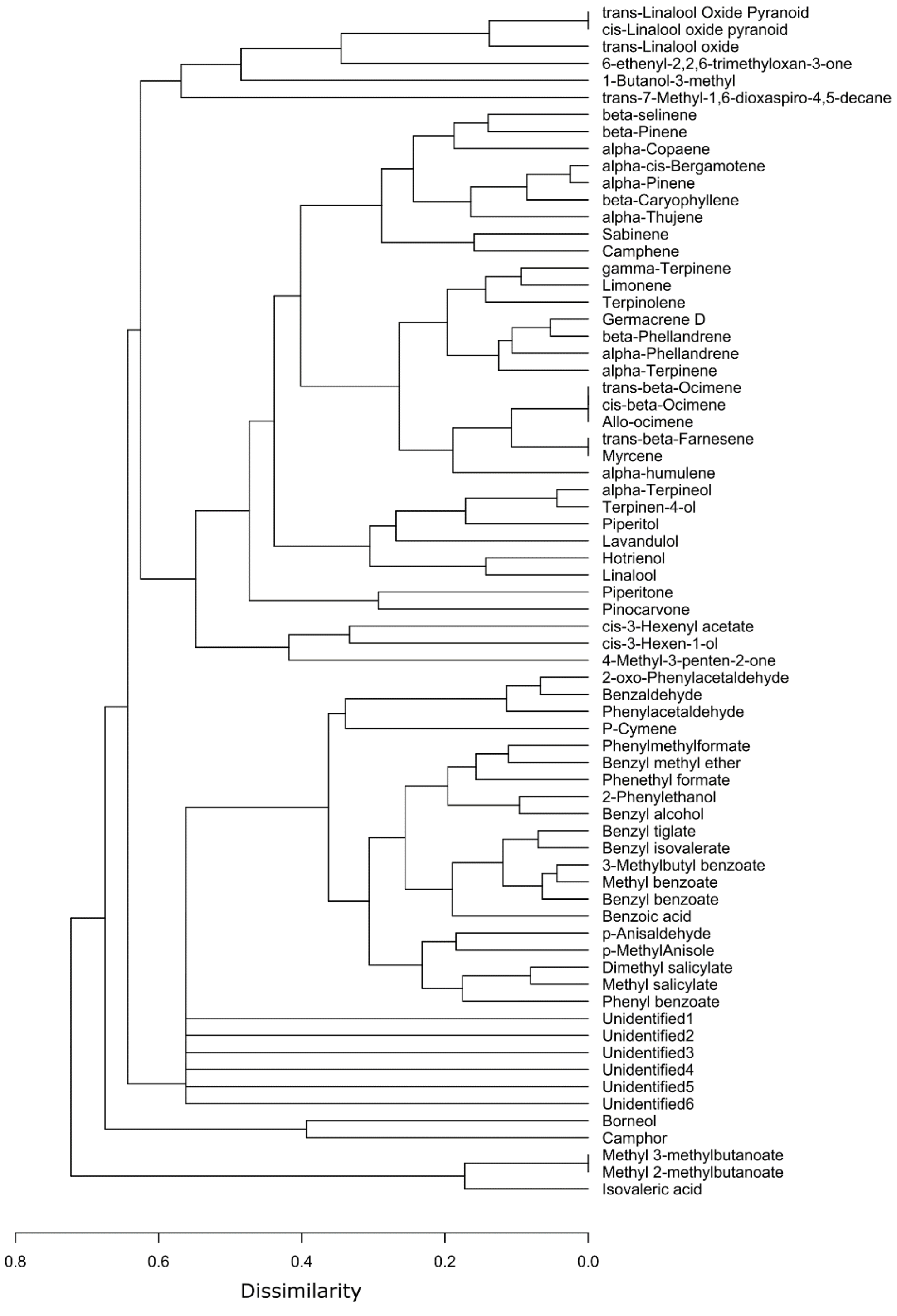

Fig. S5 Dendrogram showing compound dissimilarities for floral scent compounds found in *Achillea millefolium* and *Cirsium arvense*. This is the same dendrogram as in Fig. 4a, but with all compound names included. Data from Laure *et al.* 2016.

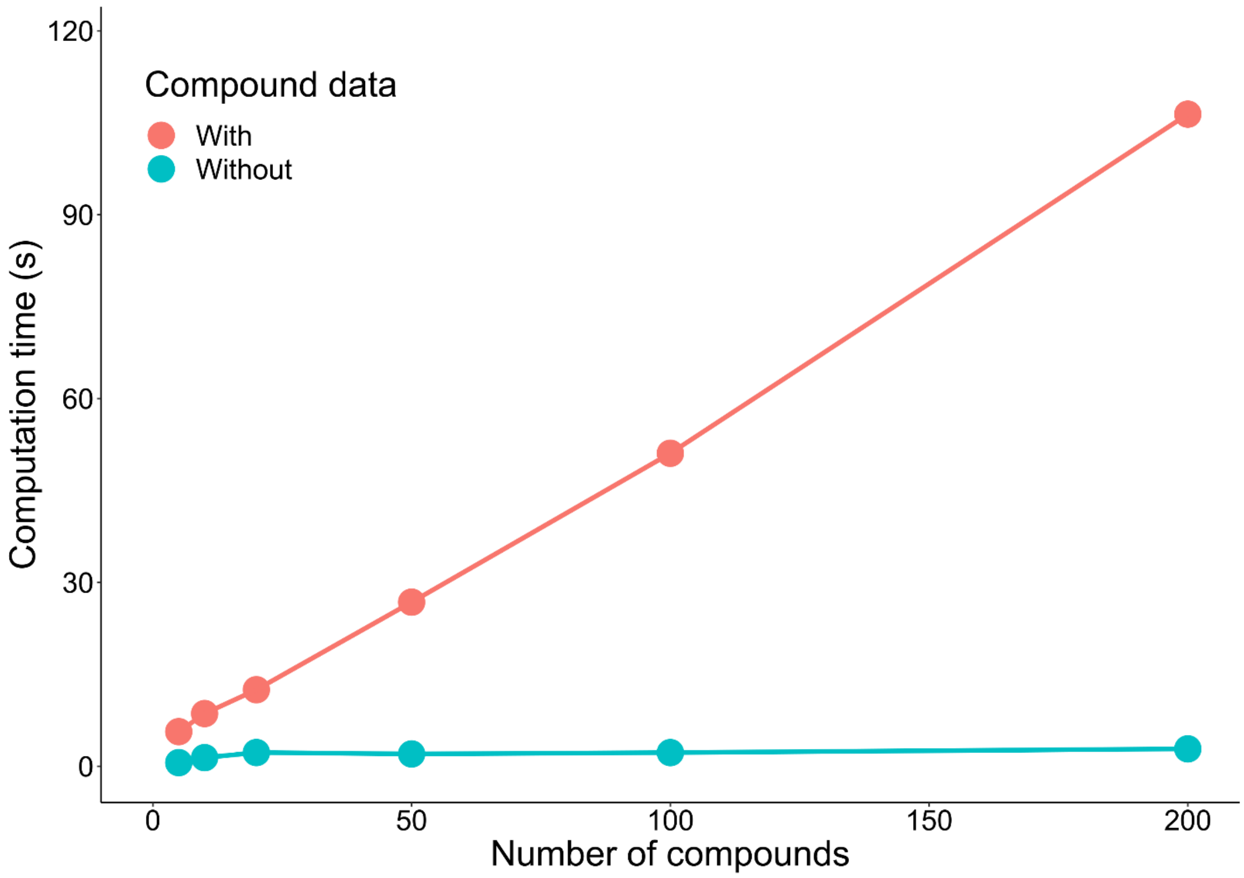

Fig. S6 Computation times for the *quickChemoDiv* function for datasets consisting of 16 samples with 5, 10, 20, 50, 100 or 200 phytochemical compounds, computed with and without including compound dissimilarity data. The *quickChemoDiv* function executes the other main functions in the package. Without compound data, computation times are 2.9 seconds for the dataset with 200 compounds. With compound data, computation times reach 106 seconds for the same dataset. This is mainly due to time taken to download data on compound structures from PubChem, which is the most time-consuming step of the analyses in this case. Comparisons were made in R 4.2.1 on a Windows 10 computer with 16 GB RAM and an Intel Core i7-7600 2.80 GHz CPU.

References

Larue A-AC, Raguso RA, Junker RR. 2016. Experimental manipulation of floral scent bouquets restructures flower-visitor interactions in the field. *Journal of Animal Ecology* **85**: 396–408.

Züst T, Strickler SR, Powell AF, Mabry ME, An H, Mirzaei M, York T, Holland CK, Kumar P, Erb M, et al. 2020. Independent evolution of ancestral and novel defenses in a genus of toxic plants (*Erysimum*, Brassicaceae). *eLife* **9**: e51712.