Supporting Information Appendix S1. Supporting Methods

The many dimensions of phytochemical diversity: linking theory to practice

William C. Wetzel[1,2] and Susan R. Whitehead[3]

[1]Department of Entomology, Michigan State University, East Lansing, MI, USA
[2]Ecology, Evolutionary Biology, and Behavior Program, Michigan State University, East Lansing, MI, USA
[3]Department of Biological Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Author emails: wcwetzel@msu.edu, swhitehead@vt.edu

Contents

I. Statistical methods for calculating diversity profiles (Fig. 3)

Diversity profile plots are based on a numbers equivalents approach to quantifying diversity (Hill 1973; Jost 2006, 2007; Jost et al. 2010) and were first applied to phenotypic and phytochemical data by Marion et al. (2015). Following Marion et al. (2015), We calculated effective diversity D of order q for K chemical compounds as

$$^qD = \left(\sum_{i=1}^{K} p_i^q \right)^{1/(1-q)},$$

where $q \neq 1$ and p_i is the proportional amount of compound i. For $q = 1$, we used

$$^1D = \exp\left(-\sum_{i=1}^{K} p_i \, \log(p_i) \right).$$

We used the following R code to calculate effective diversity D for a range of q from 0–5 for N plant samples of `n.spp` species or populations. We ran this on a matrix `dat` of proportional chemical abundance data with K columns representing each chemical compound and N rows. Our script averages across plant samples within types of plants; we did this for plotting Fig. 3, but averaging would not be necessary for a full analysis.

```
n.spp = length(unique(species)) # num species
spp = as.character(unique(species)) # species list

qa = matrix(-10, nrow=51, ncol=n.spp+1) # matrix to fill
qa[,1] = seq(0, 5, by = 0.1) # first col is q values


for(s in 1:n.spp) {

        temp = dat[species == spp[s],]

        for(i in 1:nrow(qa)) {

                q = qa[i, 1]

                # Calculate effective diversity for each plant within species
                if(q!=1) {
                Ds = apply(temp, 1, function(x) sum(x[x>0]^q)^(1 / (1-q)))
                }

                if(q==1) {
                Ds = apply(temp, 1, function(x) exp(-1 * sum(x[x>0] * log(x[x>0]))))
                }

                qa[i, s+1] = mean(Ds) # Average across all plants within species
        }
}
```

II. Statistical methods for chemical rarefaction (Fig. 4)

We used sample-based rarefaction to calculate the expected and 80% probability region of the number of compounds encountered at a given number of samples. These could represent different numbers of organs within a plant individual (Fig. 4a), different numbers of plant individuals within a population (Fig. 4b,c), different numbers of plant populations within a species (Fig. 4d), different number of plant species within a genus (Fig. 4e), different numbers of plant species within a community (Fig. 4f), or any other set of plant types within a whole, depending on the research question. We used the function `rarep` (below) to complete the rarefaction calculations. The function `rarep` takes as input a matrix with N rows, one for each sample, and K columns, one for each compound. It then randomly samples n rows from the matrix and lumps them by summing across columns of the sampled rows. Then it calculates the compound richness in the lumping of the n rows. It repeats this randomization and calculation for a number of replicates, set by the argument reps. This happens for integers $n = 1 \rightarrow$ `maxSamples`, an argument in the function that sets an upper limit for the range of n. Finally, for each n the function calculates average richness, the standard deviation of richness, and the 10th and 90th percentiles of richness. The function requires three packages: parallel (R Core Team 2018), doParallel (Microsoft and Weston 2018), and vegan (Oksanen et al. 2018). We reduced computation time by parallelizing `rarep` to spread computations across multiple cores, but it is currently written to do so only on Unix-like machines. We provide the script for a non-parallelized version, `rare`, that should work on any machine at the bottom of this section after the script for `rarep`. We would like to point out that this method currently does not use information on compounds abundances, just presence-absence. This technique, however, could be developed to include information on compound abundances such that it explores how chemistry varies quantitatively with scale.

Usage

```
rarep(mat, reps = 50, maxSamples = NA,
      numCores = detectCores() - 1)
```

Arguments

`mat`
> A matrix of compound presence or abundance with one column for each compound and one row for each sample.

`reps`
> The number of times to replicate calculations at each scale.

`maxSamples`
> The maximum number of samples to include in calculations. Default (`NA`) sets this to number of rows in `mat`.

`numCores`

The number of computing cores to use for calculations. Defaults to number detected − 1.

Value

The function `rarep` returns a data frame with five columns:

`samples`
        The number of samples averaged.

`rich`
        The mean richness across replicates.

`richS`
        The standard deviation of richnesses across replicates.

`richLower10`
        The 10th percentile of the distribution of richnesses across replicates.

`richUpper90`
        The 90th percentile of the distribution of richnesses across replicates.

Function `rarep`

```
rarep = function(mat, reps=50, maxSamples = NA, numCores = detectCores() - 1) {

        N = nrow(mat)

        if(is.na(maxSamples) | maxSamples > nrow(mat))
                MS = nrow(mat) else MS = maxSamples

        myCluster <- makeCluster(numCores, type = "FORK")
        registerDoParallel(myCluster)

        out = foreach(n = 1:MS, .combine='rbind', .verbose = FALSE) %dopar% {
                # Sample units
                temp = replicate(reps, sample(N, n, replace = FALSE))
                if(n == 1) lumpedSamples = mat[temp,] else {
                        lumpedSamples = t(apply(temp, 2, function(z) colSums(mat[z,]) ))
                }

                # Calculate richness
                richnesses = specnumber(lumpedSamples)

                # Save means and CIs across reps
                rich = mean(richnesses)
                richS = sd(richnesses, na.rm=TRUE)
                richLower10 = quantile(richnesses, probs=0.10, na.rm=TRUE)
                richUpper90 = quantile(richnesses, probs=0.90, na.rm=TRUE)

                # Combine means and CIs
                c(n, rich, richS, richLower10, richUpper90)
        }
```

```
        stopCluster(myCluster)
        out = data.frame(out)
        names(out) = c('samples', 'rich', 'richS', 'richLower10', 'richUpper90')
        return(out)
}
```

Function `rare` (non-parallel, will work on any system, only requires package `vegan`)

```
rare = function(mat, reps=50, maxSamples = NA) {

        N = nrow(mat)

        if(is.na(maxSamples) | maxSamples > nrow(mat))
                MS = nrow(mat) else MS = maxSamples

        dat = data.frame(samples=1:MS, rich=NA, richS=NA, richLower10=NA,
        richUpper90=NA)

        pb = txtProgressBar(min = 0, max = MS, initial = 0, style=3)

        for(n in 1:MS){
                # Sample units
                temp = replicate(reps, sample(N, n, replace = FALSE))
                if(n == 1) lumpedSamples = mat[temp,] else {
                        lumpedSamples = t(apply(temp, 2, function(z) colSums(mat[z,]) ))
                }

                # Calculate richness
                richnesses = specnumber(lumpedSamples)

                # Save means and CIs across reps
                dat[n,'rich'] = mean(richnesses)
                dat[n,'richS'] = sd(richnesses, na.rm=TRUE)
                dat[n,'richLower10'] = quantile(richnesses, probs=0.10, na.rm=TRUE)
                dat[n,'richUpper90'] = quantile(richnesses, probs=0.90, na.rm=TRUE)

                setTxtProgressBar(pb, n)
        }
        return(dat)
}
```

4

III. Statistical methods for spatial chemical rarefaction (Fig. 5)

The aim of this analysis is to explore how phytochemical diversity changes with increasing spatial scale, taking into account the spatial locations of each plant unit. This is important because spatial autocorrelation in chemistry has the potential to influence the amount of diversity experienced by organisms interacting with plants. This is true as long as those organisms are limited in their movements relative to the extent of a study area. We demonstrate this phenomenon by using a spatial rarefaction method based on nearest neighbors, calculated using the function `rareSpat` (below). This function calculates richness across each sample and its 0–`maxNeighbors` nearest neighbors. Richness for a sample and 0 neighbors is simply the richness in each sample. Richness for a sample and its 1st closest neighbor is the richness aggregated across the sample and its neighbor. The function requires package `spatstat` (Baddeley & Turner 2005, Baddeley et al. 2015) for identifying nearest neighbors. We would like to point out that this method fits well when questions focus on how perceived chemistry varies with the number of adjacent discrete plant units. For example, this would be relevant when considering insects that visit different numbers of plants. An alternative way organisms interact spatially with plants is by encountering plants within an area of movement. Our method could be easily adapted to show how perceived chemistry varies with area (rather than with number of neighbors) by calculating richness aggregated across all samples within r distance of random spatial points. This distance could be varied across a biologically relevant range, and the compound richness could be plotted as a function of radius r or area.

Usage

```
rareSpat(mat, spat, maxNeighbors = 10)
```

Arguments

`mat`
> A matrix of compound presence or abundance with one column for each compound and one row for each sample.

`spat`
> A two-column matrix of spatial coordinates.

`maxNeighbors`
> The maximum number of neighboring samples to include in calculations. Defaults to 10.

Value

The function `rareSpat` returns a matrix with one column for each plant sample and one row for each spatial scale, in terms of number of plants. This means the first row gives the richness within each plant (no neighbors). The second row is the richness across each plant and its nearest neighbor. The third row is the richness across each plant and its nearest two neighbors and so forth up to `maxNeighbors` + 1.

## Function `rareSpat`

```
rareSpat = function(mat, spat, maxNeighbors=10) {

        N = nrow(mat)
        mN = ifelse(maxNeighbors > (N-1), N-1, maxNeighbors)
        rich = matrix(-11, nrow=mN+1, ncol=N) # -11 is a placeholder

        # Identify nearest neighbors up to mN
        nns = nnwhich(spat, k=1:mN)

        # Calculate richness for each focal plant n and its i nearest neighbors
        for(n in 1:N){
                for(i in 0:mN){
                        matTemp = mat[ c(n, nns[n , 0:i]) , ] # n b/c need focal plant
                        rich[i+1,n] = ifelse(is.matrix(matTemp),
                                sum(colSums(matTemp[]) > 0),
                                sum(matTemp > 0)
                        )
                }
        }
        return(rich)
}
```

IV. The relationship between sample concentration and detected compound richness

We illustrated the analytical relationship between extract concentration and observed chemical richness by diluting a single leaf extract of ground cherry (*Physalis angulata*) to ten concentrations and assessing richness as the number of peaks detected using GC-MS. Our results may not be surprising to those who have analyzed various concentrations of complex plant samples for methods optimization, but they provide clear evidence that the richness detected in a sample is strongly and positively dependent on the total phytochemical abundance (Fig. S1). In the main text, we explain the significance of this phenomenon for the study of phytochemical diversity.

For this illustration, we prepared an extract of *Physalis angulata* leaves for analysis via gas chromatography/mass spectrometry (GC-MS). Plant material was obtained from a greenhouse population at Virginia Tech grown from seed that was originally collected from a wild population outside Raleigh, North Carolina, U.S.A. Leaves were collected in bulk from 2-3 plants, immediately frozen, stored at -80°C, and then lyophilized prior to analysis. Dry leaves (2.25g) were ground to a fine powder in 50mL centrifuge tubes via a FastPrep-24 (MP Biomedicals) and extracted two times in methanol for 30min on a shaker tray at room temperature with a ratio of 20mL MeOH per gram of plant material. The combined extracts were evaporated to dryness under a stream of ultra-pure nitrogen. To remove non-target polar metabolites, extracts were re-suspended in 1:1 water:hexanes at a ratio of 100mL solvent per gram of crude extract. Extracts were hand-shaken for 60 sec, centrifuged, and the hexanes layer (containing withanolides and other non-polar metabolites) was transferred to a clean tube and evaporated to dryness. The extract was then re-suspended in dichloromethane at a ratio of 1mL per 10mg extract. This solution was then serially diluted by 50:50 ratios with dichloromethane to 11 concentrations ranging from 10mg/ml to 0.0098 mg/ml. Samples were analyzed on an Agilent 7820GC coupled with a 5977 MS set to 70eV electron ionization and equipped with an HP5-MS column (30 m x 0.25 mm i.d., 0.25μm film thickness; Agilent Technologies, Santa Clara, CA, USA). Ultra-pure helium was used as a carrier gas at a flow rate of 1mL/min. We used a splitless injection volume of 1μl with an injector temperature of 280°C. The temperature program was as follows: initial temperature 120°C, initial hold time 2 min; ramp 1: 15°C min^{-1} to 250°C, hold time 0 min; ramp 2: 5°C min^{-1} to 300°C, hold time 10 min; total run time of 30.67 min. Data were recorded as TIC chromatograms and processed using MSD ChemStation software. We used automated integration parameters and counted the number of detected peaks in each sample as an estimate of phytochemical richness.
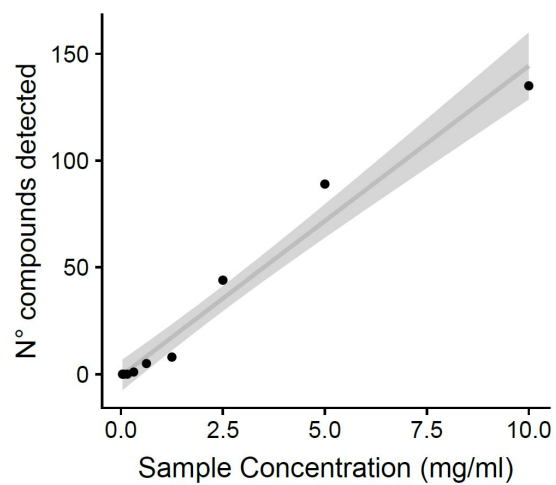
Fig. S1. The number of metabolites detected by GC-MS analysis in ground cherry (*Physalis angulata*) as a function of the sample concentration (mg extract / ml solvent). Note that these samples all come from a serial dilution of a single homogenized sample.

V. Exploration of diversity–abundance relationship in *Physalis angulata*

The overall abundance of phytochemicals often varies among samples within a study (e.g., among plant genotypes or species, or between herbivore induced and undamaged plant individuals). Given the abundance-dependence of observations of phytochemical diversity (Fig. 6, Fig. S1), within-study variation in abundance could bias estimates of phytochemical diversity. We propose that studies, when feasible, examine diversity per unit phytochemical biomass or use other methods that allow diversity to be disentangled from total abundance. This would allow us to be sure we are not confounding the effects of diversity and abundance and enable rigorous tests of potential biological explanations for the abundance-dependence of phytochemical diversity. Here we present an initial exploration of two potential methods for correcting phytochemical richness estimates to account for variation in overall phytochemical abundance among samples. However, we acknowledge that more work is needed to test these methods, that these methods would be just one part of a study's careful methods development phase, and that these methods decisions should be made only in the context of the specific system being studied.

We explored two potential methods for correcting diversity estimates for abundance using an analysis of withanolides in nine samples of *Physalis angulata* (see chemical methods below). In these samples, as in other systems (Fig. 6), observed richness was positively correlated with abundance, which we estimated both based on gravimetric measurements of extract mass and on summed peak areas (Fig. S2a-b). This suggests that, for at least some biological questions, we may want to use a metric for diversity that accounts for variation in abundance. A temptingly simple way to assess richness per unit abundance would be to simply divide richness by abundance. This, however, yields a potentially spurious negative correlation between corrected richness and abundance owing to their mathematical dependence (Figs. S2c-d). Instead, we suggest measuring all sample extract masses gravimetrically (reporting extract masses is standard in natural products chemistry but often overlooked in ecology) and diluting extracts for analysis at a fixed ratio of solvent to extract mass. We did this by transferring aliquots of the plant extracts used in Fig. S2a to pre-weighed vials, evaporating to dryness, and re-weighing to obtain the total extract mass. Samples were then re-suspended in solvent at a constant ratio of extract mass to solvent (0.5 mg/100uL) prior to HPLC analysis.

The ideal ratio for analysis will vary across systems and should ideally be determined by serially diluting extracts to various concentrations (as in Fig. S1) and choosing a concentration that maximizes observable phytochemical diversity while maintaining adequate peak resolution. Researchers will have to subjectively assess this in their system, and the optimal concentration will vary across instruments and methods. For example, in our *Physalis angulata* studies, we found that HPLC analysis at a concentration of 0.5mg/ml resulted in a detected phytochemical richness of 46-55 compounds across samples (Fig. S2). Yet, in GC/MS analysis of the same sample type, we detected only 5 unique compounds at a similar concentration (Fig. S1). This emphasizes that phytochemical richness cannot be interpreted outside of the analytical context in which it was measured. By analyzing samples at multiple concentrations at the start of the study, researchers can gain an appreciation for the variation in detected phytochemical richness and choose a concentration that best represents the variation present in their system.

9

In our *Physalis* study, once samples were diluted at a set ratio of extract mass to solvent, our analysis indicated no correlation between abundance and diversity (Fig. S2e), suggesting that at least a portion of the abundance-diversity relationship is explained by the analytical relationship between sample concentration and observed richness. If samples cannot be dried and total phytochemical abundance must be estimated based on peak areas, an alternative solution is to use the residuals of the diversity–abundance relationship (e.g., regression in Fig. S2b), which provides relative estimates of high or low richness of samples given their phytochemical abundances (Fig. S2f). However, if it is possible to evaporate an extraction to dryness to obtain the extract mass, we would suggest that a better solution may be to differentially dilute samples according to a set ratio of extract mass to solvent and estimate richness directly from chromatograms (as in Fig. S2e). This would allow a more straightforward interpretation of richness as a direct estimate of the number of compounds per unit total phytochemical abundance.

While these methods have limitations and may not fit every study, they have the potential to reveal important new patterns in the ecology and evolution of phytochemical diversity. At a minimum, method development in phytochemistry should involve running samples at multiple concentrations to understand how abundance affects diversity estimates. We need to account for the intimate links between diversity and abundance otherwise we risk confounding their effects.


Chemical methods—We prepared a set of extracts from leaves collected from nine individuals of *Physalis angulata* following methods in Cao et al. (2015). Plant material was collected from a greenhouse population, immediately frozen in liquid nitrogen and lyophilized. Dry material (100 mg) was ground in 2mL microcentrifuge tubes via a FastPrep-24 (MP Biomedicals) with stainless steel beads. To each tube, we added 500µl of 1:1 dichloromethane:methanol and sonicated for 45 min. Samples were then centrifuged and two separate 100µl aliquots were transferred to pre-weighed micro-inserts for HPLC vials. Samples were evaporated to dryness under a stream of ultra-pure nitrogen, and the inserts were re-weighed to obtain the crude extract masses. Note that these gravimetric estimates of extract mass were subject to measurement error. The masses obtained from our two theoretically identical extracts varied by up to 33% (Fig. S2e compared to a or c). This is not surprising considering that the extract masses (0.57-1.61mg) were only a small percentage of the mass of the vial (approximately 550mg), and nearing the limit of precision of our balance. Additional error could have arisen if samples are not well-homogenized prior to removing aliquots. The masses of our two aliquots were strongly correlated ($R^2=0.83$), indicating that our gravimetric measurements provide a reasonable estimate of extract mass, but ideally extract masses should be obtained on a high-precision balance and/or scaled up to decrease proportional measurement error. Researchers should carefully consider sources of error in their own methods and work to minimize this variation as much as possible.

Once extract masses were obtained, one set of aliquots was re-suspended in constant volume of 100µl and a second set was re-suspended using a constant concentration of 0.57 mg extract per 100µl solvent. This concentration was determined based on the minimal solvent volume that needed to be added to the sample with the smallest mass to allow analysis. All other samples

were diluted to maintain equal ratios of extract mass to solvent across samples. Samples were vortexed and analyzed using an Agilent 1260 high-pressure liquid chromatograph (HPLC) equipped with a diode array detector and a Synchronis aQ C18 reversed-phase column (250 x 4.6mm and 5μm particle size; Thermo-Fisher Product #97305-254630). Monitoring was at 220nm for optimal detection of withanolides (Cao et al. 2015). The injection volume was 10μl and the solvent system was water (A) and acetonitrile (B) with a linear gradient (30-80% B; 0-18min) followed by an isocratic elution (80% B; 18-25min) at a constant flow rate of 1ml min^{-1}. Data were processed using ChemStation software. Individual compounds were not identified; however, comparisons of UV absorbance spectra with authentic standards suggests most compounds were withanolides. To obtain a relative estimate of total withanolide abundance across samples, we summed the peak areas of all detected peaks. To obtain estimates of phytochemical richness, we counted the number of total peaks detected using automated integration.
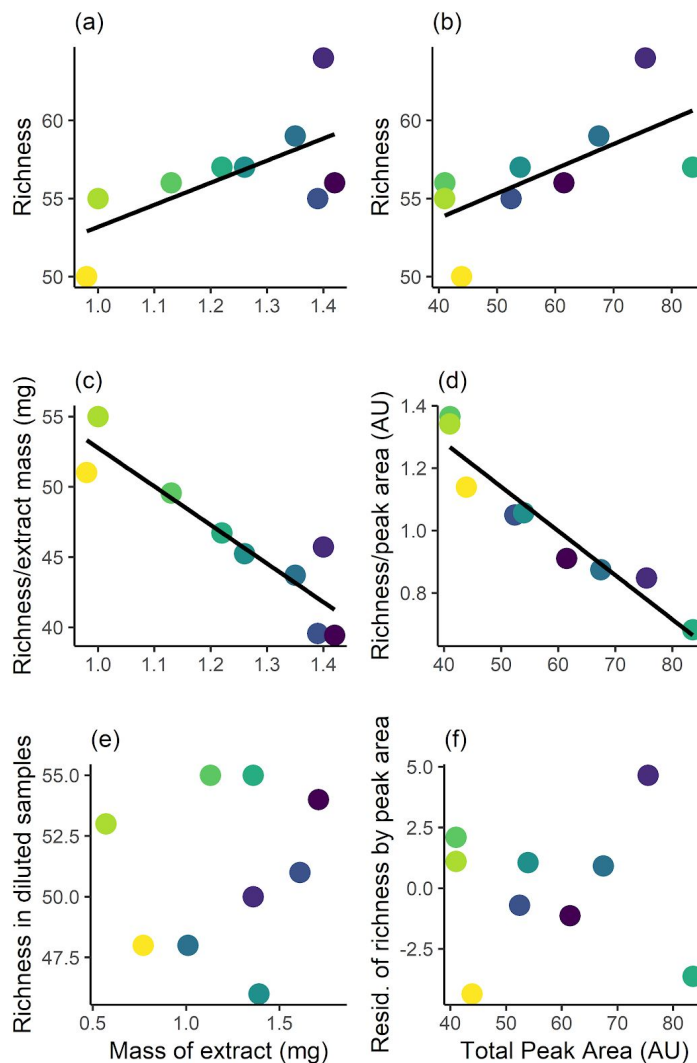
Fig. S2. An exploration of potential methods that estimate phytochemical richness and are not biased by the analytical relationship between richness and abundance. Data show withanolide richness and abundance in the leaves of nine *Physalis angulata* individuals (each individual represented by a single color consistent across panels). For panels a-d & f, extracts were prepared using typical methods, in which a consistent mass of dry plant material is extracted, dried, and re-suspended in a set volume of solvent for analysis. Richness and abundance were then calculated in various ways from the same analysis. For panel e, second aliquots of the same extracts were transferred to a separate vial, dried, and re-suspended using a set ratio of dry extract mass to solvent volume. Variation in extract masses between e and other panels is due to variation between aliquots and/or measurement error. Regardless of whether abundance is estimated as phytochemical extract mass (a, c, e) or as total peak area (b, d, f), there is a strong correlation between richness and abundance (a, b) and a potentially spurious negative correlation when richness is simply divided by abundance (c, d). To obtain an estimate of richness that is independent of abundance, we suggest diluting samples proportional to extract mass (e). If this is not possible, an alternative is to estimate relative richness across samples using the residuals of the relationship between total peak area and number of compounds (b), which would also provide an estimate of relative richness that is independent of abundance (f).
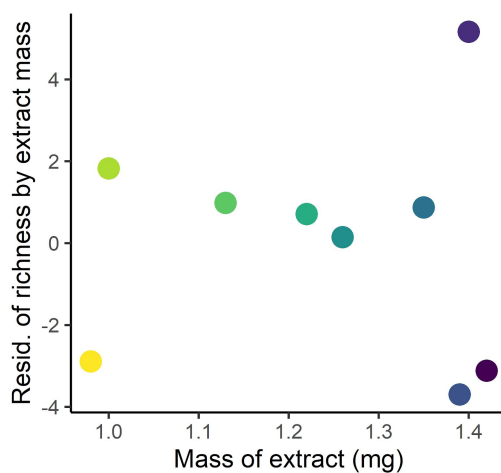
Fig. S3. Residuals of relationship between the mass of phytochemical extract (mg) and phytochemical richness in the *Physalis angulata* correction study (i.e., residuals from the relationship shown in Fig. S2a). The residuals of this relationship could provide another relative estimate of compound richness across samples that is independent of abundance. However, if it is possible to evaporate an extraction to dryness to obtain the extract mass, we would suggest that a better solution may be to differentially dilute samples according to a set ratio of extract mass to solvent and estimate richness directly from chromatograms (as in Fig. S2f). This would allow a more straightforward interpretation of richness as a direct estimate of the number of compounds per unit total phytochemical abundance.

VI. References

Baddeley, A., Rubak, E., & Turner, R. (2015). Spatial Point Patterns: Methodology and Applications with R. London: Chapman and Hall/CRC Press.

Baddeley, A., & Turner, R. (2005). spatstat: An R package for analyzing spatial point patterns. *J. Stat. Soft.*, **12**, 1-42.

Cao, C.-M., Kindscher, K., Gallagher, R.J., Zhang, H., Timmermann, B.N. (2015). Analysis of major withanolides in *Physalis longifolia* Nutt. by HPLC-PDA. *J. Chromatogr. Sci.*, **53**, 1044-1047.

Hill, M.O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, **54**, 427-432.

Jost, L. (2006). Entropy and diversity. *Oikos*, **113**, 363-375.

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, **88**, 2427-2439.

Jost, L., Devries, P., Walla, T., Greeney, H., Chao, A., & Ricotta, C. (2010). Partitioning diversity for conservation analyses. *Divers. Distrib.*, **16**, 65-76.

Marion, Z.H., Fordyce, J.A & Fitzpatrick, B.M. (2015). Extending the concept of diversity partitioning to characterize phenotypic complexity. *Am. Nat.*, **186**, 348-361.

Microsoft Corporation & Weston, S. (2018). doParallel: Foreach Parallel Adaptor for the 'parallel' Package. R package version 1.0.14. https://CRAN.R-project.org/package=doParallel.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., & Wagner, H. (2018). Vegan: Community Ecology Package. R package version 2.5-2. https://CRAN.R-project.org/package=vegan.

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.