Estimating belowground plant abundance with DNA metabarcoding

Silvia Matesanz[1*], David S. Pescador[1], Beatriz Pías[2], Ana M. Sánchez[1], Julia Chacón-Labella[3], Angela Illuminati[1], Marcelino de la Cruz[1], Jesús López-Angulo[1], Adrián Escudero[1]

[1] Área de Biodiversidad y Conservación, Universidad Rey Juan Carlos. Tulipán s/n, 28933, Móstoles, Spain

[2] Departamento de Biodiversidad, Ecología y Evolución. Universidad Complutense de Madrid. José Antonio Nováis 2, 28040, Madrid, Spain.

[3] Departamento de Medio Ambiente. Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA) Ctra. de la Coruña Km 7.5, 28040 Madrid, Spain

* Author for correspondence: silvia.matesanzgarcia@gmail.com

Running title: Estimating root biomass with metabarcoding

Abstract

Most work on plant community ecology has been performed aboveground, neglecting the processes that occur in the soil. DNA metabarcoding, where multiple species are computationally identified in bulk samples, can help overcome the logistical limitations involved in sampling plant communities belowground. A major limitation of this methodology is, however, the quantification of species' abundances based on the percentage of sequences assigned to each taxon. Using root tissues of the five dominant species in a semiarid Mediterranean shrubland (*Bupleurum fruticescens, Helianthemum cinereum, Linum suffruticosum, Stipa pennata* and *Thymus vulgaris*), we built pairwise mixtures of relative abundance (20, 50 and 80% biomass), and implemented two methods (linear models fits and correction indices) to improve root biomass estimates. We validated both methods with multispecies mixtures that simulate field-collected samples. For all species, we found a positive and highly significant relationship between the percentage of sequences and biomass in the mixtures (R^2 = 0.44-0.66), but the equations for each species (slope and intercept) differed among them, and two species were consistently over- and under-estimated. The correction indices greatly improved the estimates of biomass percentage for all five species in the multispecies mixtures, and reduced the overall error from 17% to 6%. Our results show that, through the use of post-sequencing quantification methods on mock communities, DNA metabarcoding can be effectively used to determine not only species' presence but also their relative abundance in field samples of root mixtures. Importantly, knowledge on these aspects will allow to study key, yet poorly understood, belowground processes.

Keywords: DNA metabarcoding, plant abundance, root biomass, sequence, Mediterranean shrubland, coexistence, mock communities, *rbcL* region

1. INTRODUCTION

A critical question in plant ecology is how communities are structured in space and time. In this still-unresolved debate, community ecologists attempt to disentangle the relative role of key stochastic and deterministic processes, such as niche differentiation, biotic interactions, and environmental filtering to determine plant species coexistence (Chase & Leibold 2003; Götzenberger et al. 2012; Gravel et al. 2006; HilleRisLambers et al. 2012; Vellend 2010). A major limitation is the fact that our understanding on the structure of plant diversity stems from data collected almost entirely aboveground. However, a large proportion of the community biomass can be located belowground (Hilbert & Canadell 1995; Poorter et al. 2012; Schenk & Jackson 2002), particularly in stressful habitats, and as such, both plant-soil and plant-plant interactions may have important implications for community-level processes (Bardgett et al. 2014; Bever et al. 2010; Casper & Jackson 1997; Philippot et al. 2013; Wardle et al. 2004).

The main constraint to sampling plant communities belowground is that reliable species identification in natural conditions based solely on morphological root traits is extremely difficult or simply unfeasible in many cases (Silva & Rego 2003). In this context, the development of molecular methods such as DNA metabarcoding, spurred by the emergence of next-generation sequencing, has had a significant impact on biodiversity assessments (Schuster 2007; Taberlet et al. 2012). DNA metabarcoding involves the simultaneous identification of multiple species based on the amplification and sequencing of a common target DNA region from an environmental or community bulk sample (Deiner et al. 2017; Hollingsworth et al. 2009; Kress et al. 2005; Taberlet et al. 2012). For instance, for plant communities, DNA metabarcoding has been successfully used to recreate their current and past composition from soil-derived DNA (Fahner et al. 2016; Jørgensen et al. 2012; Porter et al. 2016; Yoccoz et al. 2012) or to

identify the floral composition of honey (Hawkins et al. 2015). Going belowground, a few studies have also assessed the richness and composition of temperate and tropical plant communities using root mixtures or individual root fragments (Hiiesalu et al. 2012; Jones et al. 2011; Kesanakurti et al. 2011).

There is mounting evidence that DNA metabarcoding is a robust method to assess biodiversity. Indeed, some studies even found higher DNA-based diversity compared to traditional sampling methods (reviewed in Deiner et al. 2017). However, there is currently an intense debate on the use of read number to quantify DNA metabarcoding results, with some authors limiting its use to strictly detect occurrence, whilst others advocate a quantitative approach (see discussion on e.g. Bell et al. 2019; Deiner et al. 2017; Fonseca 2018; Porter & Hajibabaei 2018). Ideally, the percentage of sequences assigned to each taxa during DNA metabarcoding would closely reflect the species' abundance (biomass, number of individuals, etc.) in the bulk sample. Building on this simple assumption, several studies have attempted the direct use of the observed percentage of DNA sequences to estimate species' abundances in communities of microbes (Amend et al. 2010), stoneflies (Elbrecht & Leese 2015), fish and amphibians (Evans et al. 2016; Pont *et al.* 2018), zooplankton (Harvey et al. 2017) and fungi (Merges et al. 2018). However, many factors operating during DNA extraction, amplification and sequencing as well as the inherently compositional nature of the data can alter the correspondence between the percentage of reads retrieved and the species' abundance (Cristescu 2014; Deiner et al. 2017; Elbrecht & Leese 2015; Gloor, Macklaim, Pawlowsky-Glahn, & Egozcue, 2017; Pawlak et al. 2015; Polz & Cavanaugh 1998; Porter & Hajibabaei 2018). Indeed, studies where such correspondence is lacking suggest that the use of uncorrected, observed percentages may render strongly biased estimates of abundance (see e.g. Bell et al. 2019; Deagle et al. 2013; Lim et al. 2016),

and effort is now being devoted to the improvement of quantification methods (Thomas et al. 2016 and references therein; McLaren et al. 2019; Nichols et al. 2018; Piñol et al. 2015). In this context, the use of mock communities, i.e. a defined mixture of tissues with known species composition and relative abundance (biomass), can be a useful tool to improve biomass estimates in DNA metabarcoding studies (see e.g. Thomas *et al*. 2016 for a comprehensive example using prey fish mixtures).

Quantification of species' biomass through DNA metabarcoding can be critical in the study of belowground community structure. Compared to other plant communities, Mediterranean shrublands are highly diverse, and up to 80% biomass can appear belowground (Hilbert & Canadell 1995). In these water-limited systems, belowground plant-plant interactions can be equally important, or even more, than those occurring aboveground (Casper & Jackson 1997). However, experimental evidence on their direction, strength and correspondence to the interactions occurring aboveground is scarce (but see Armas & Pugnaire 2011). Furthermore, because species' abundances are markedly heterogeneous and leptokurtic, with a few very abundant species and many rare ones (Chacón-Labella et al. 2017; Chacón-Labella et al. 2016; McGill et al. 2007), presence-absence data fails to accuretaly reflect the structure of the plant community. Therefore, in order to gain insights on the mechanisms that determine plant community structure and to build a global coexistence theory, we should expand our focus belowground and compare these patterns to those aboveground. To do this, we need robust information not only on the presence of species in the soil but also on their relative abundance across space.

In this study, we built mock communities with varying composition and abundance of five selected species from Mediterranean shrublands, and used a DNA metabarcoding approach on these root mixtures, to move beyond species detection and

115 estimate species' relative biomass. We implement two post-sequencing quantification
116 methods. First, we fit linear models to assess whether the percentage of reads (DNA
117 sequences) can be used to robustly estimate percentage of root biomass, and second, we
118 compute correction indices that control for potential biases and improve the relationship
119 between sequences and biomass percentages (see Thomas et al. 2016). In addition, to
120 determine the possibility to apply our results to field-collected samples, we validate
121 both methods with multispecies realistic samples. To our knowledge, this is the first
122 study aimed at the improvement of a quantitative DNA metabarcoding approach in
123 plants.
124

125 **2. MATERIALS AND METHODS**

126 **2.1 Plant community and species selection**

127 The study plant community is a species-rich semiarid Mediterranean shrubland
128 established in limestone and gypsum soil in the central Iberian Peninsula. Perennial
129 cover ranges from 40 to 60%, and is mainly dominated by small chamaephytes and
130 grasses. It is a highly diverse community, with around 50 perennial species found at the
131 local scale (e.g.≈8000 individuals from 48 species in 60 m^2; Chacón-Labella *et al.*
132 2016). The distribution of individuals across species is highly heterogeneous, with a few
133 species accounting for a high proportion of the total number of individuals. Given the
134 disproportionate influence of the most abundant species, we selected the five most
135 common species in the community for our study (Fig. 1): *Thymus vulgaris* L.
136 (Lamiaceae), *Helianthemum cinereum* (Cav.) Pers. (Cistaceae), *Linum suffruticosum* L.
137 (Linaceae), *Bupleurum fruticescens* L. (Apiaceae), and *Stipa pennata* L. (Poaceae). The
138 selected species have different phylogenetic origins, life forms, and can account for as
139 much as 65% of the total number of individuals in the community (data not shown).

2.2 Sampling material and creation of root mixture mock communities

We collected root samples in the shrublands near Orusco de Tajuña (Madrid, Spain, 40°17'17.5"N 3°12'19.4"W). For each selected species, we uprooted 5-10 adult individuals with unequivocal taxonomic identification. All individuals were collected within 24h, bagged separately, stored in a cooler and immediately transferred to the lab at Universidad Rey Juan Carlos. Upon arrival, their root system was thoroughly washed and separated from the soil, and roots from all individuals of the same species were pooled and maintained in cool water until sample preparation. We created mock communities (hereafter mixtures) based on mixtures of root biomass, varying both the species composition, richness and the percentage of biomass of each species in each sample. Note that the use of root mixtures (community DNA) rather than DNA extracted from soil samples (environmental DNA *sensu* Deiner et al. 2017) allowed to quantify biomass of actively growing plants, avoiding the presence of persistent DNA from long-dead individuals (Baird & Hajibabaei, 2012).

The communities were created by cutting small pieces of roots (removed of excess water by patting them with paper towel) and weighing them in a Mettler Toledo MX5microbalance (1 µg precision; Mettler Toledo, Columbus, OH, USA) the same day of collection in the field. All mock communities contained 100 mg of fresh root biomass, and were immediately frozen at -80ºC for later DNA metabarcoding analyses. We created two different types of mock communities:

1) Pairwise mixtures, with two species present in different proportions (20:80, 50:50, or 80:20; all pairwise combinations with three replicates per type of community, N = 90 samples; Fig. 1). The pairwise mixtures were used to: i) determine the match between the percentage of biomass and the percentage of DNA sequences (hereafter

reads) obtained via linear model fits, and ii) calculate the correction indices (see section on statistical analyses).

2) Multispecies mixtures, with the five selected species. We first combined them at the same proportion (20:20:20:20:20; one mixture with three replicates, N = 3) and then we created communities where the percentage of one species (either *Helianthemum cinereum* or *Stipa pennata*) was progressively increased and that of the other four was maintained equal (10.0:22.5:22.5:22.5:22.5, 40:15:15:15:15, 60:10:10:10:10, and 80:5:5:5:5; eight types of mixtures with three replicates, N = 24). These two species were chosen because they had shown either relatively lower or higher amplification in a previous pilot study (data not shown). The multispecies mixtures were used to validate the calculated linear fit parameters and correction indices. See Table S1 for details on the composition of each type of mixture.

2.3 DNA metabarcoding on root mixtures mock communities

DNA was extracted from each mixture (and four isolation blanks) in the lab at Universidad Rey Juan Carlos using the DNEasy Plant Minikit (Qiagen, CA, USA) and shipped to the AllGenetics laboratories (AllGenetics & Biology SL, A Coruña, Spain). For library preparation, we amplified a fragment of the *rbcL* chloroplast gene (550 bp) using primers rbcLa-F (5' ATGTCACCACAAACAGAGACTAAAGC3'; Levin *et al.* 2003 and rbcLa-R (5' GTAAAATCAAGTCCACCRCG 3'; Kress *et al.* 2009), to which the Illumina sequencing primer sequences were attached at the 5' ends. We selected the *rbcL* region because it has repeatedly been shown to be a robust barcode for plants (Hollingsworth et al. 2009; Kress et al. 2009), and because it allowed the taxonomic identification of most members of the entire study community at the species level. A series of two PCRs were carried out, the first to amplify the selected region and the

second to attach the index sequences required for multiplexing different libraries in the same sequencing pool. PCRs were carried out in a final volume of 25 µl, containing 2.5 µl of template DNA, 0.5 µM of the primers, 12.5 µl of Supreme NZYTaq 2x Green Master Mix (NZYTech, Lisboa, Portugal), and ultrapure water up to 25 µl. The reaction mixture consisted of an initial denaturation at 95 ºC for 5 min, followed by 30 cycles of 95 ºC for 30 s, 52 ºC for 30 s, 72 ºC for 30 s, and a final extension step at 72 ºC for 10 minutes. The second PCR had identical conditions but only 5 cycles and 60 ºC as the annealing temperature. Two negative controls with no DNA were included to check for contamination during library preparation. Portugal), The libraries were run on agarose gels stained with GreenSafe (NZYTech, Lisboa, and their size visualized under UV light. They were then purified using the Mag-Bind RXNPure Plus magnetic beads (Omega Biotek, GA, USA), pooled in equimolar amounts and sequenced in a run of the MiSeq PE300 (Illumina, CA, USA).

The quality of the Illumina paired-end raw data was checked using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc), and the raw reads were quality-filtered using Geneious 11.1.2 (www.geneious.com). The PCR primers were removed and a region at the 3' end of each file was trimmed according to the average Phred score (minimum Phred quality score of 20). Since the amplicons were too long to allow for the overlap of forward and reverse reads, R1 and R2 reads were concatenated using the fuse.sh script implemented in the BBmap package (Bushnell 2014). The sequences were labelled (demultiplexed) using the script multiple split libraries.py implemented in Qiime (Caporaso et al. 2010). The label is added to the headers of the FASTQ file in order to identify each sample when sequences are combined to perform downstream analysis. The resulting FASTA file was processed using the VSEARCH bioinformatic tool (Rognes et al. 2016). Sequences were dereplicated (-derep fullength), clustered at a

similarity threshold of 100 % (-cluster fast,--centroids option), and sorted (-sortbysize). Artifacts (such as point mutations and chimeras) that may be generated during PCR and sequencing were filtered during the bioinformatic pipeline. *De novo* chimera detection was carried out using the UCHIME algorithm (Edgar et al. 2011) implemented in VSEARCH.

The taxonomic identification was performed using an in-house constructed reference database containing the representative *rbcL* sequences (553bp) of 45 species from 18 families of plants from the study community that had been collected in the same study site and individually sequenced (see sequences in XX). Since the query sequences mapped only to the 5' and 3' ends of the reference sequences, the central region of the reference sequences was previously deleted to perfectly match the query sequences, resulting in a final length of 517bp (see a similar approach in Vizcaíno *et al*. 2018). The taxonomic identification was performed by querying the clustered representative sequences against our reference database using the -usearch global option of VSEARCH with a 99% similarity threshold. Finally, the script mesas-uc2clust.py was used to obtain an OTU table listing the number of sequences from each OTU found in each sample. Based on the results of this table, a quality-filtering was applied to remove the OTUs with a number of sequences lower than 0.005% of the total number of sequences (Bokulich et al. 2013; Edgar 2013).

In DNA metabarcoding studies, it has been observed that a low percentage of the reads of a library can be assigned to another library. This phenomenon, known as mistagging or index jumping is the result of the misassignment of the indices during library preparation, sequencing, and/or demultiplexing steps (Esling, Lejzerowicz & Pawlowski 2015; Bartram et al. 2016). To correct for this, the low abundance OTUs of each sample (0.1% threshold) were removed. Finally, only the OTUs that matched any

reference sequence in the database at a similarity of 99% were kept in the OTU table. The unidentified OTUs were removed from the OTU table for downstream analysis. These unidentified OTUs only accounted for an average 0.90% of the total reads before filtering.

2.4 Statistical analyses

All the OTUs assigned to the same species were combined before analysis. For each sample, we calculated the percentage of DNA reads assigned to each species, as the number of reads for each species divided by the total number of reads in the sample. To check whether the percentage of reads reflected the species-specific percentage of biomass in each mixture, and to improve our inference ability in the cases where there was not a reliable match between both aspects, we used two different methods: 1) linear model fits and 2) creation of correction indices (relative correction factors *sensu* Thomas et al. 2016).

1. *Linear model fits and match to the identity function*: for each species, we used the pairwise mixtures where it was present (N = 36) to compute the best linear fit ($y = ax + b$ in intercept-slope form, where a is the slope of the line and b is the intercept) between the percentage of retrieved reads (y) and the percentage of biomass (x; 20, 50 or 80%), using the *lm* function in R (R Core Team 2017). In order for the percentage of reads to be used directly as an estimate of the species' biomass percentage in a sample, the equation obtained for a given species would have to closely match the identity function, i.e. a linear equation where the intercept is not significantly different from zero and the slope is not significantly different from one ($y = x$). If this is the case, the percentage of reads found after sequencing and filtering could be directly used to estimate the original percentage of biomass in the sample. Therefore, we used the

parameter estimates from the equations fitted to the data of each species to verify the significance of the test on the intercept (b = 0), and performed a two-tailed t-test to compare the slope of the fit to a slope of one (a = 1)(see a similar approach on Diaz-Real et al. 2015).

To assess whether the inclusion of other species affected the relationship between the percentage of biomass of each species in the samples and the percentage of retrieved reads, we fitted linear equations to the data including the multispecies mixtures (N = 63 for each species). Again, we tested whether the intercept was different from zero and whether the slope of the equation was significantly different from one. We finally compared the slope of both fits (only pairwise mixtures vs. all mixtures), performing an analysis of covariance (ANCOVA) for each species with type of community (pairwise vs. all) as predictor, percentage of biomass as covariate and percentage of reads as dependent variable. A significant interaction between the predictor and the covariate indicates that the slope fitted to the different communities is not the same. Different intercepts and/or slopes in both equations would indicate that the fits are affected by the species richness and composition of the samples, and therefore that employing the percentage of sequences of DNA as estimation of biomass in communities of different richness would be severely biased, even for the same species.

2. *Correction indices:* for each species, we calculated a percentage-specific correction index using the pairwise mixtures, based on the percentage of biomass in the sample and the percentage of reads retrieved (relative correction factors presented in Thomas et al. 2016). Specifically, the correction index for a species A at percentage p was calculated as:

$$Correction\ index_{A_p} = \frac{Biom\ sp_{A_p}}{100 - Biom\ sp_{A_p}} * \frac{\sum_{i \neq A}^{s} \sum_{j=1}^{r} Reads\ sp_{ij}}{\sum_{i \neq A}^{s} \sum_{j=1}^{r} Reads\ sp_{Aij}}$$

Where r is the number of replicates for each type of community (i.e., $r = 3$), s is the number of species considered (i.e., $s = 5$), $Biom\ sp_{A_p}$ is the percentage of biomass of the species A, $Reads\ sp_{Aij}$ is the number of reads obtained for species A in combination with species i in each replicate j, and $Reads\ sp_{ij}$ is the number of reads obtained for each of the other species i in replicate j. A correction index was computed for each species and percentage by averaging the number of reads in all the mixtures where it was combined with the all the other species at a specific percentage. For instance, the 20% correction index for *Thymus vulgaris* was computed using all the samples where the percentage of biomass of *Thymus vulgaris* was 20% and that of the second species (either one of the other four) was 80%. When the correction index is≈1, the percentage of reads retrieved robustly reflects the percentage of biomass on the sample. When the correction index is well above or below 1, the percentage of reads obtained is underestimated or overestimated compared to the percentage of biomass, respectively. We computed a 20, 50, 80% correction index and an average index (average of the three) for each species.

The correction indices computed for each species were used to recalculate the number of reads of each species in the multispecies mixtures (by multiplying the index by the retrieved number of reads), and subsequently transform these numbers to percentage of reads. This was performed for all indices of each species (20, 50, 80% and average). Then, these corrected percentages were compared to the actual percentage of biomass of each species in the multispecies mixtures. To assess the correction ability of the computed indices, we calculated the error for each type of community (averaging the three replicates). For each species and percentage of biomass, we computed the average absolute difference between the actual percentage of biomass and the corrected percentage of reads. We also computed the average absolute differences for the original,

uncorrected estimates. To assess the effect of the correction and whether it was similar for all species, we used a two-way ANOVA, with correction (corrected vs. uncorrected) and species as predictors, and absolute error as the dependent variable. This was followed by species-specific one-way ANOVAs to test the effect of the correction on each species individually. Lower average error in the corrected percentages would indicate that the correction indices based on the pairwise mixtures could improve the estimates in field-collected samples of different species richness and composition.

An alternative method to compute correction indices is the use of a control species (see details in Thomas et al. 2016). In this approach, the correction index for each target species is computed using only the pairwise mixtures where the target and the control species are present. We computed species-specific correction indices using the 50% pairwise mixtures with three different control species: *L. suffruticosum, H. cinereum* and *S. pennata*. To assess whether the use of control-based correction indices also improved the percentages of biomass, we again used these indices to correct the percentage of reads of each species in the multispecies mixtures. Finally, for each species and multispecies mixture, we computed the average error (as defined above) of the corrections using the different control-based indices.

3. RESULTS

3.1 Performance of DNA metabarcoding with root mixture mock communities

For most samples (> 94%), DNA metabarcoding successfully recreated the species composition of the mixtures (mock communities), i.e. all the species added to a mixture were found during sequencing. Only for a few samples where the percentage of biomass of *Stipa pennata* was low ($\leq$ 20%), no sequences for this species were recovered (Supporting Information Table S1). Five species from the study plant community that were not added to the root mixtures were also detected in a few samples (N = 16), but

were always found at low percentages (range: 0.13-15.06%, average: 2.65%; see Table S1 for details on retrieved compositions and percentage of reads of each species and sample). Specifically, in more than 90% of the samples, more than 95% of the retrieved sequences were assigned exclusively to the species added to each mixture (Supporting Information Fig. S1).

3.2 Evaluation of linear model fits and match to the identity function

For the five selected species, the linear models fitted to the data were highly significant and had a positive slope, i.e. the percentage of reads of each species increased when the percentage of biomass of the species in the sample also increased (adjusted R^2 = 0.32-0.50; Fig. 2, Supporting Information Table S2). However, both the slope and the intercept of the fitted models differed among species. Specifically, the data obtained from the pairwise mixtures of *B. fruticescens*, *L. suffruticosum* and *T. vulgaris* could be fitted to a linear equation with an intercept not significantly different from zero (b = 0, Table S2). Similarly, the t-tests showed that the slopes of the lines for these species were not significantly different from one (P > 0.34 in all cases; Table S2). This indicates that the percentage of reads retrieved for these three species may be used directly to estimate the percentage of biomass on the samples. Conversely, the lines fitted to *H. cinereum* and *S. pennata* had intercepts significantly different from zero (significantly higher/lower than zero for *H. cinereum* and *S. pennata*, respectively; Table S2). Similarly, the slopes were significantly different from one for both species (P = 0.02 and P < 0.001 for *H. cinereum* and *S. pennata,* respectively), indicating that the percentage of reads was consistently higher (*H. cinereum*) or lower (*S. pennata*) than the percentage of biomass in the sample.

363 The models including all mixtures showed similar results to those fitted only
364 with the pairwise mixtures (Table S2). Indeed, for all species except *S. pennata*, the
365 slopes of both equations (only pairwise vs. all mixtures) were not significantly different
366 (not significant interaction 'percentage of biomass × type of community' in ANCOVA).
367 For *B. fruticescens*, *L. suffruticosum* and *T. vulgaris*, the intercept and the slope of the
368 lines fitted to all the mixtures were not significantly different from zero and one,
369 respectively, matching again the identity function, and the fit improved for all species
370 (adjusted R^2 = 0.56-0.66; Fig. 2, Table S2). This indicates that species richness and
371 composition did not significantly alter the fits for these three species. Conversely, when
372 the multispecies mixtures were added to the data of the remaining species (*H. cinereum*
373 and *S. pennata*), the equations had again intercepts significantly different from zero and
374 slopes significantly higher/lower than one (P < 0.001for both *H. cinereum* and *S.*
375 *pennata*; Fig. 2).
376
377 **3.3 Evaluation of correction indices**
378 We found a wide variation among species for the correction indices computed with all
379 the pairwise mixtures and a control species (Fig. 3 and Supporting Information Table
380 S3). Using all pairwise mixtures, the correction indices for *B. fruticescens, L.*
381 *suffruticosum* and *T. vulgaris* were close to one, and slightly increased when the
382 percentage of biomass in the sample increased. However, for the remaining species, the
383 correction indices were much lower (*H. cinereum*) or much higher (*S. pennata*) than one
384 (Fig. 3, lower panels), indicating a consistent overestimation and underestimation of the
385 percentage of reads compared to biomass percentages.
386 When the number of reads of the multispecies mixtures were recalculated using
387 the 50% correction indices, the recalculated percentage of reads closely matched the

388 actual biomass percentage in the multispecies samples (Fig. 4), and the average absolute
389 error (absolute difference between the actual percentage of biomass and the percentages
390 of reads) was significantly reduced (Fig. 5, significant differences in the average
391 absolute error between corrected and uncorrected percentage of reads, $P < 0.0001$). This
392 error reduction was not equal among species (significant 'species × correction'
393 interaction, $P < 0.001$), and was especially relevant for *H. cinereum* and *S. pennata*,
394 where their overestimation and underestimation in the uncorrected percentage of reads
395 were significantly improved when the correction indices were applied (e.g. Fig. 4 c and
396 d). For these species, the error between reads and biomass was significantly reduced
397 after correction (Fig. 5). For *B. fruticescens*, *L. suffruticosum* and *T. vulgaris*, the
398 recalculation of reads with the correction indices also improved the match between the
399 percentage of reads and biomass (e.g. Fig. 4 g and i), although the reduction of the error
400 was not significant for these species (Fig. 5). The use of the 50%, 80% and average
401 correction indices rendered very similar results (from ≈17% error in the uncorrected
402 samples to ≈6% in the corrected percentages), but the correction of the proportions was
403 lower when the 20% correction indices were used (≈9% overall error in the corrected
404 percentages).

405 Similarly, the use of a control species to calculate the correction indices
406 improved the estimation of percentages of biomass in most cases (Supporting
407 Information Table S4), but the correction of the proportions varied depending on the
408 choice of control species (Table S4), and the error reduction was on average lower than
409 when all pairwise mixtures were used to calculate the correction indices (Supporting
410 Information Fig. 2).

411

412 **4. DISCUSSION**

413 Our study provides a straightforward and simple protocol to overcome one of the main
414 shortcomings in DNA metabarcoding studies, the estimation of species' relative
415 abundance based on the percentage of DNA sequences (reads) recovered. Through the
416 use of purposefully-designed root mock communities, we test the efficacy of two
417 complementary and easy-to-implement methods and provide robust estimates of plant
418 biomass percentages in realistic multispecies samples. This is, to the best of our
419 knowledge, the first study to validate a quantitative DNA metabarcoding in plant
420 communities using root mixtures.

421 The use of metabarcoding is revolutionizing plant ecology studies, since
422 detection of the so-called hidden diversity provides new insights to open questions in
423 this field (see e.g. Yoccoz et al. 2012). However, the possibility of using DNA
424 metabarcoding results to estimate species' abundances has been a subject of debate
425 since the onset of this methodology. Due to reported inconsistencies in past attempts,
426 recent revisions suggest that a conservative approach may be to treat metabarcoding
427 results as presence-absence data (Deiner et al. 2017; Porter & Hajibabaei 2018).
428 However, our study suggests that accurate quantification of species roots' biomass may
429 be robustly done, provided that previous quantification studies using mock communities
430 with target species are performed. Importantly, results from mock communities of root
431 mixtures (both fitted models and corrected read percentages) can be then safely used to
432 robustly estimate root biomass in field-collected samples, since estimated biomass
433 percentages (using both methods) were not significantly altered by species composition,
434 richness and species' relative abundance in the samples.

435 Our results have important implications for plant community ecology.
436 Understanding how and to what extent stochastic and deterministic processes determine
437 plant coexistence and community assemblages in plant communities remains an

438 unresolved question, despite the intense research effort devoted to this topic over the
439 last decades (Götzenberger et al. 2012; Gravel et al. 2006; HilleRisLambers et al. 2012).
440 A few authors have recognized that part of this knowledge gap could be filled if we
441 complement our current framework, mainly based on characterization of aboveground
442 processes, expanding our focus belowground (Bever et al. 2010; Wardle et al. 2004). In
443 this context, DNA metabarcoding has successfully been used in a few instances to
444 describe patterns of species richness and its distribution belowground (Hiiesalu et al.
445 2012; Jones et al. 2011; Kesanakurti et al. 2011), but quantification attempts were
446 lacking. The prospect of using DNA metabarcoding on root mixtures to detect not only
447 the presence of species but also to estimate species' abundances constitutes a step
448 further towards a deeper understanding of plant coexistence and community
449 assemblages, especially at the fine scales where roots interact. Knowledge on the
450 patterns of root biomass distribution will provide insights on the correspondence
451 between above- and belowground distributions, plant-plant interactions and plant-soil
452 feedbacks (Brandt et al. 2013; Kulmatiski et al. 2008).

453 An ideal scenario for DNA metabarcoding studies would be that the proportion
454 of DNA sequences obtained after high-throughput sequencing closely reflected the
455 percentage of biomass of each species in the bulk sample, irrespective of the sample
456 composition and the relative occurrence of each species. If this was true for our plant
457 community, the percentage of DNA sequences assigned to each species could be readily
458 used to estimate the percentage of root biomass in field-collected samples. For the five
459 study species, we indeed found a positive and highly significant relationship between
460 the percentage of biomass in the pairwise mixtures and the percentage of reads
461 recovered for each species (Fig. 2). However, the parameters of the statistical
462 relationship (slope and intercept) widely varied among species, and for two of them (*H.*

463 *cinereum* and *S. pennata*), the best fit rendered biased estimates of biomass percentages,
464 despite the observed positive correlation. For instance, for *H. cinereum*, the estimated
465 percentage of biomass using the fitted equation for the pairwise communities with 20,
466 50 or 80% biomass rendered 60, 75 and 90% biomass estimates, respectively, due to the
467 high intercept of the fitted line (the opposite, i.e. a sharp underestimation of biomass
468 proportions, occurred for *S. pennata*). These results point out that a significant positive
469 relationship between percentage of biomass and percentage of reads is not sufficient to
470 transform presence-absence data into quantitative estimates (despite its current wide
471 use). To robustly achieve the latter, the line fitted for a given species would need to be
472 statistically equivalent to the identity function. For three of our study species (*T.*
473 *vulgaris*, *B. fruticescens* and *L. suffruticosum*), we found such equivalence between
474 biomass and reads percentages. This match was not altered when the data from the
475 pairwise mixtures was combined to the multispecies samples, suggesting that, at least
476 for these three species, the relationship between root biomass and reads percentages is
477 maintained regardless of the number of species (two versus five) and the species'
478 biomass percentage (from 5 to 80%). However, even when the fit is equivalent to the
479 identity function, the predicted abundance estimated by the linear model may be poor
480 (e.g. fitting with a large residual error). Overall, our results call for caution on the direct
481 use of sequence percentages to approximate relative biomass or abundance based on the
482 existence of a positive relationship between both aspects (see e.g. Elbrecht & Leese
483 2015; Hiiesalu et al. 2012; Pont et al. 2018) or on the mere assumption that such
484 relationship exists (see e.g. Merges et al. 2018), and highlight the need to test the
485 properties (statistical parameters) rather than just the existence of a significant
486 relationship between percentage of reads and abundance for each species individually.

Our second approach involved the use of species-specific correction indices (based on the relative correction factors recently proposed by Thomas et al. 2016) obtained from either all the pairwise mixtures or using a control species, which were then used to correct the percentage of sequences in the multispecies samples. Importantly, these recalculated read percentages generally improved the match between the percentage of reads and the actual biomass in the multispecies mock communities (Fig. 4 and Table S4), and reduced the overall error compared to the uncorrected percentages (Figs. 5 and S2). The best results, i.e. the lowest error, was obtained when the percentage of reads were recalculated using the indices computed with all pairwise mixtures, as they closely mirrored the biomass percentages in each multispecies mock community (Fig. 4). This indicates that the use of such correction indices represents a successful way to obtain quantitative estimates in plant DNA metabarcoding studies. Several pieces of evidence support this claim. First, reliable estimates of biomass percentages were obtained for all five species after adjusting the percentages of reads, which suggests that this method can be generalized to other species in the community. This was the case even for the two species that significantly deviated from the identity function due to consistent over- and underestimation (*H. cinereum* and *S. pennata*). Indeed, the bias reduction –calculated as the difference between sequence and actual biomass percentages– between the observed and corrected percentages was higher for these two species (note that the lower bias reduction in the other three species was due to the fact that their correction indices were in all cases very close to one, i.e. no strong deviation between biomass and uncorrected reads percentages). Second, the correction index calculated for each species was computed based on pairwise mixtures of different compositions (each species combined with the other four) and then applied to multispecies samples, which highlights that these indices are robust to changes both in

512 species richness and composition. And third, the indices calculated from pairwise
513 mixtures where the species were found at different proportions remained relatively
514 constant (only when percentages were low, i.e. 20%, did the indices substantially
515 differed; Fig. 3), and efficiently corrected samples where biomass percentages varied
516 widely. This indicates that these correction indices are also relatively robust to varying
517 species' biomass percentages. Our results concur with those by Thomas et al. (2016),
518 the only other existing implementing correction indices, who found that control-based
519 correction of reads proportion greatly improved relative abundances in fish mixtures.

520 This study also allowed to compare the correction ability of differently-
521 computed indices. Although those based on all pairwise mixtures outperformed
522 correction indices based on the use of a control species, the latter also resulted in
523 improved biomass estimates compared to uncorrected ones. The use of a control species
524 to compute correction indices has the advantage that the number of pairwise mixtures
525 needed is significantly lower (e.g. in a five-species study, only four pairwise mixtures
526 are needed if the fifth species is the control, but 10 pairwise mixtures are needed to
527 compute indices from all pairwise mixtures), which can significantly reduce the
528 complexity and cost of the study. However, the reduction of error widely varied
529 depending on the choice of control species (Fig. S2), which introduces a source of
530 uncertainty since the control species needs to be chosen a priori. In practice, the
531 decision of how to compute correction indices will depend on a variety of factors,
532 including the species richness of the study community, existing knowledge of the
533 performance of species during metabarcoding, etc.

534 Our study also helped to validate the effectiveness of DNA metabarcoding using
535 the *rbcL* region for the simultaneous identification and quantification of multiple taxa in
536 root mixtures from Mediterranean shrublands. For most samples, the species that

formed each mock community were successfully recovered during sequencing. For a few samples, however, our approach recovered species –either the study species or other species from our plant community– that had not been included in those specific mixtures, although in general they accounted for a very small percentage of the DNA sequences in each sample (Fig. S1). These infrequent mismatches between the created (prepared root mixtures) and recreated (after sequencing) species composition can be due to species cross-contamination during root sampling and mock community preparation, or due to mistagging (i.e. index/tag jumping) during the DNA metabarcoding pipeline (Coissac et al. 2012; Schnell et al. 2015). Importantly, they help to identify aspects for improvement in metabarcoding studies (Deiner et al. 2017; Porter & Hajibabaei 2018). Furthermore, it is worth to note that the choice of the appropriate barcode may depend on the type of plant community and the source of DNA samples (community DNA, environmental DNA, etc.). Future studies should also incorporate several markers to determine the consistency of the correction indices across different barcodes (Hollingsworth et al. 2009).

In conclusion, we propose that the use of mock communities varying in species composition and biomass structure may be a useful first step for the reliable quantification of DNA metabarcoding results in other plant communities, implementing a combined approach where linear fits and correction indices are used. However, the substantial differences observed among the study species –both in the linear fits and the correction indices– indicates that quantification methods need to be applied on a species-level basis. Different sources of bias may occur during DNA extraction (e.g. differential DNA concentration per tissue biomass across species; see also Haling et al. 2011) or PCR amplification (e.g. differential primer specificity; Cristescu 2014; Deiner et al. 2017; Elbrecht & Leese 2015; Pawluczyk et al. 2015; Porter & Hajibabaei 2018),

562 leading to some species being consistently under- or over-estimated during sequencing.
563 Therefore, it is highly unlikely that biomass percentages can be estimated for all species
564 in a community using the same linear fit or correction index. Furthermore, the ability to
565 perform quantitative DNA metabarcoding will largely depend on the number of species
566 in the study community, which in turn determines the amount of mock communities
567 needed to implement corrections. In this context, prior knowledge on the species
568 composition of the community (i.e. the existence of a robust reference library) and the
569 selection of study species (e.g. dominant, keystone species) are critical for the
570 successful implementation of reliable quantification methods. Finally, our results also
571 suggest that the indiscriminate use of uncorrected percentages of sequences as a proxy
572 for species' biomass without previous quantification tests such as the one presented here
573 may render strongly biased results for many species.

574

575 **ACKNOWLEDGEMENTS**

576 We are in debt to Joaquín Vierna, Neus Marí-Mena and Antón Vizcaino (AllGenetics)
577 for their dedication and support during DNA metabarcoding and bioinformatic analysis.
578 We are also grateful to Carlos Díaz (URJC) for his help in the collection of plant
579 material. We are also grateful to three anonymous reviewers and editor for their
580 thorough revision of our manuscript. Funding was provided by MINECO grant ROOTS
581 (CGL2015-66809-P).

582

583 **AUTHOR CONTRIBUTION**

584 AE, DSP and SM conceived and designed the study. DSP, BP and SM prepared the
585 mock communities. JC-B prepared the reference database. AI, MC, DSP and SM

analyzed the data. All authors contributed to the discussion and interpretation of the results. SM wrote the manuscript, with input from all other authors.

DATA ACCESSIBILITY

Data has been deposited in Dryad (doi:10.5061/dryad.dm4t39t).

REFERENCES

Amend AS, Seifert KA, Bruns TD (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology* **19**, 5555-5565.

Armas C, Pugnaire FI (2011) Belowground zone of influence in a tussock grass species. *Acta Oecologica-International Journal of Ecology* **37**, 284-289.

Baird, D. J., & Hajibabaei, M. (2012). Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. Molecular Ecology, 21(8), 2039-2044.

Bardgett RD, Mommer L, De Vries FT (2014) Going underground: root traits as drivers of ecosystem processes. *Trends in Ecology & Evolution* **29**, 692-699.

Bell KL, Burgess KS, Botsch JC, et al. (2019) Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. Molecular Ecology **28**, 431-455.

Bever JD, Dickie IA, Facelli E, *et al.* (2010) Rooting theories of plant community ecology in microbial interactions. *Trends in Ecology & Evolution* **25**, 468-478.

Bokulich NA, Subramanian S, Faith JJ, *et al.* (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods* **10**, 57.

Brandt AJ, Kroon H, Reynolds HL, Burns JH (2013) Soil heterogeneity generated by plant–soil feedbacks has implications for species recruitment and coexistence. *Journal of Ecology* **101**, 277-286.

Bushnell, B. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7,** 335.

Casper BB, Jackson RB (1997) Plant competition underground. *Annual Review of Ecology and Systematics* **28**, 545-570.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology* **21**, 1834-1847.

Cristescu ME (2014) From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution* **29**, 566-571.

Chacón-Labella J, Cruz M, Escudero A (2017) Evidence for a stochastic geometry of biodiversity: the effects of species abundance, richness and intraspecific clustering. *Journal of Ecology* **105**, 382-390.

Chacón-Labella J, de la Cruz M, Escudero A (2016) Beyond the classical nurse species effect: diversity assembly in a Mediterranean semi-arid dwarf shrubland. *Journal of Vegetation Science* **27**, 80-88.

Chase JM, Leibold MA (2003) *Ecological niches: linking classical and contemporary approaches* University of Chicago Press.
Deagle BE, Thomas AC, Shaffer AK, Trites AW, Jarman SN (2013) Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? *Molecular Ecology Resources* **13**, 620-633.
Deiner K, Bik HM, Mächler E, *et al.* (2017) Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Molecular Ecology* **26**, 5872-5895.
Diaz-Real J, Serrano D, Piriz A, Jovani R (2015) NGS metabarcoding proves successful for quantitative assessment of symbiont abundance: the case of feather mites on birds. Experimental and Applied Acarology **67**, 209-218.
Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10**, 996.
Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 27(16), 2194-2200.
Elbrecht V, Leese F (2015) Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. *PLoS ONE* **10**, e0130324.
Evans NT, Olds BP, Renshaw MA, *et al.* (2016) Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources* **16**, 29-41.
Fahner NA, Shokralla S, Baird DJ, Hajibabaei M (2016) Large-Scale Monitoring of Plants through Environmental DNA Metabarcoding of Soil: Recovery, Resolution, and Annotation of Four DNA Markers. PLoS ONE 11, e0157505.
Fonseca VG (2018) "Pitfalls in relative abundance estimation using eDNA metabarcoding". *Molecular Ecology Resources* **18**, 923-926.
Götzenberger L, de Bello F, Bråthen KA, *et al.* (2012) Ecological assembly rules in plant communities—approaches, patterns and prospects. *Biological Reviews* **87**, 111-127.
Gravel D, Canham CD, Beaudet M, Messier C (2006) Reconciling niche and neutrality: the continuum hypothesis. *Ecology Letters* **9**, 399-409.
Haling RE, Simpson RJ, McKay AC, et al. (2011) Direct measurement of roots in soil for single and mixed species using a quantitative DNA-based method. Plant and Soil **348**, 123-137.
Harvey JB, Johnson SB, Fisher JL, Peterson WT, Vrijenhoek RC (2017) Comparison of morphological and next generation DNA sequencing methods for assessing zooplankton assemblages. *Journal of Experimental Marine Biology and Ecology* **487**, 113-126.
Hawkins J, de Vere N, Griffith A, *et al.* (2015) Using DNA metabarcoding to identify the floral composition of honey: A new tool for investigating honey bee foraging preferences. *PLoS ONE* **10**, e0134735.
Hiiesalu I, OePIK M, Metsis M, *et al.* (2012) Plant species richness belowground: higher richness and new patterns revealed by next-generation sequencing. *Molecular Ecology* **21**, 2004-2016.
Hilbert DW, Canadell J (1995) Biomass partitioning and resource allocation of plants from Mediterranean-type ecosystems: possible responses to elevated atmospheric CO2. In: *Global Change and Mediterranean-Type Ecosystems*, pp. 76-101. Springer.

HilleRisLambers J, Adler P, Harpole W, Levine J, Mayfield M (2012) Rethinking community assembly through the lens of coexistence theory. *Annual Review of Ecology, Evolution, and Systematics* **43**.

Hollingsworth PM, Forrest LL, Spouge JL, *et al.* (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* **106**, 12794-12797.

Jones FA, Erickson DL, Bernal MA, *et al.* (2011) The Roots of Diversity: Below Ground Species Richness and Rooting Distributions in a Tropical Forest Revealed by DNA Barcodes and Inverse Modeling. *PLoS ONE* **6**.

Jørgensen T, Kjaer KH, Haile J, *et al.* (2012) Islands in the ice: detecting past vegetation on Greenlandic nunataks using historical records and sedimentary ancient DNA Meta-barcoding. *Molecular Ecology* **21**, 1980-1988.

Kesanakurti PR, Fazekas AJ, Burgess KS, *et al.* (2011) Spatial patterns of plant diversity below-ground as revealed by DNA barcoding. *Molecular Ecology* **20**, 1289-1302.

Kress WJ, Erickson DL, Jones FA, *et al.* (2009) Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proceedings of the National Academy of Sciences*, pnas. 0909820106.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences* **102**, 8369-8374.

Kulmatiski A, Beard KH, Stevens JR, Cobbold SM (2008) Plant–soil feedbacks: a meta-analytical review. *Ecology Letters* **11**, 980-992.

Levin RA, Wagner WL, Hoch PC, et al. (2003) Family-level relationships of Onagraceae based on chloroplast rbcL and ndhF data. American Journal of Botany 90, 107-115.

Lim NKM, Tay YC, Srivathsan A, *et al.* (2016) Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. *Royal Society Open Science* **3**.

McGill BJ, Etienne RS, Gray JS, *et al.* (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters* **10**, 995-1015.

Merges D, Bálint M, Schmitt I, Böhning-Gaese K, Neuschulz EL (2018) Spatial patterns of pathogenic and mutualistic fungi across the elevational range of a host plant. *Journal of Ecology* **106**, 1545-1557.

Nichols RV, Vollmers C, Newsom LA, et al. (2018) Minimizing polymerase biases in metabarcoding. Molecular Ecology Resources **18**, 927-939.

Pawlak AR, Mack RN, Busch JW, Novak SJ (2015) Invasion of Bromus tectorum (L.) into California and the American Southwest: rapid, multi-directional and genetically diverse. *Biological Invasions* **17**, 287-306.

Pawluczyk M, Weiss J, Links MG, *et al.* (2015) Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. *Analytical and Bioanalytical Chemistry* **407**, 1841-1848.

Philippot L, Raaijmakers JM, Lemanceau P, van der Putten WH (2013) Going back to the roots: the microbial ecology of the rhizosphere. *Nature Reviews Microbiology* **11**, 789-799.

Piñol J, Mir G, Gomez-Polo P, Agustí N (2015) Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. Molecular Ecology Resources **15**, 819-830.

Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. Applied and Environmental Microbiology 64, 3724-3730.
Pont D, Rocle M, Valentini A, *et al.* (2018) Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. *Scientific reports* **8**, 10361.
Poorter H, Niklas KJ, Reich PB, *et al.* (2012) Biomass allocation to leaves, stems and roots: meta-analyses of interspecific variation and environmental control. *New Phytologist* **193**, 30-50.
Porter TM, Hajibabaei M (2018) Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology* **27**, 313-338.
Porter TM, Shokralla S, Baird D, Golding GB, Hajibabaei M (2016) Ribosomal DNA and Plastid Markers Used to Sample Fungal and Plant Communities from Wetland Soils Reveals Complementary Biotas. PLoS ONE 11, e0142759.
Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, e2584.
Schenk HJ, Jackson RB (2002) Rooting depths, lateral root spreads and below-ground/above-ground allometries of plants in water-limited ecosystems. *Journal of Ecology* **90**, 480-494.
Schnell IB, Bohmann K, Gilbert MTP (2015) Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources* **15**, 1289-1303.
Schuster SC (2007) Next-generation sequencing transforms today's biology. *Nature Methods* **5**, 16.
Silva JS, Rego FC (2003) Root distribution of a Mediterranean shrubland in Portugal. *Plant and Soil* **255**, 529-540.
Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* **21**, 2045-2050.
Thomas AC, Deagle BE, Eveson JP, Harsch CH, Trites AW (2016) Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources* **16**, 714-726.
Vellend M (2010) Conceptual synthesis in community ecology. *The Quarterly Review of Biology* **85**, 183-206.
Wardle DA, Bardgett RD, Klironomos JN, *et al.* (2004) Ecological linkages between aboveground and belowground biota. *Science* **304**, 1629-1633.
Yoccoz N, Bråthen K, Gielly L, *et al.* (2012) DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology* **21**, 3647-3655.

Figure 1. Workflow implemented to validate DNA metabarcoding quantification methods. 1) Selection of the most dominant species in the study plant community (based on the number of individuals). 2) Creation of the pairwise and multispecies mixtures, i.e. root mock communities of known composition and varying percentage of biomass proportion (see Table S1 for detailed information on composition of all mock communities). 3) DNA metabarcoding and bioinformatics pipeline: DNA extraction, PCR (*rbcL* gene), next-generation-sequencing and taxonomic assignment. 4) Quantification methods and validation: calculation of the percentage of reads (DNA sequences) assigned to each species in each mixture and comparison to the actual percentage of biomass in the sample via linear fits and correction indices. Validation of both methods with multispecies mixtures.

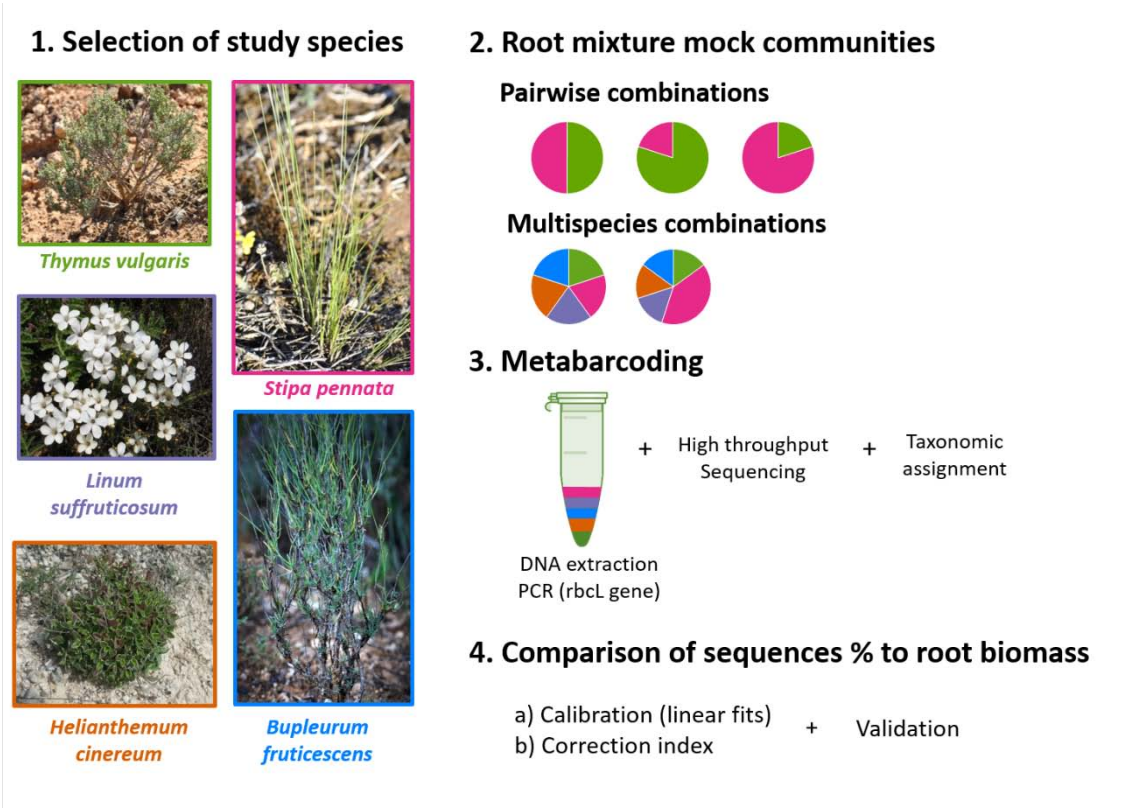


Figure 2. Percentage of DNA reads recovered as a function of the percentage of biomass in each mixture (mock community). Each panel presents all the mixtures where each species was present. Mean ± standard error of three replicates of each mixture are shown. Different colors of the symbols reflect the accompanying species in the pairwise mixtures (i.e. they correspond to colors of scientific names in each panel; e.g. pink circles in the top left panel represent pairwise mixtures of *B. fruticescens* with *S. pennata*). Grey diamond shapes represent the proportion of DNA reads obtained for each species in the multispecies mixtures. The best linear fit (colored line), adjusted R^2, intercept (*b*) and slope (*a*) estimates are also shown for each species (pairwise and multispecies samples combined). A significant intercept and/or slope indicate significant differences from zero and one, respectively. The grey dashed line represents the intercept = 0 and slope = 1 fit (identity function). *** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$; ns, not significant.

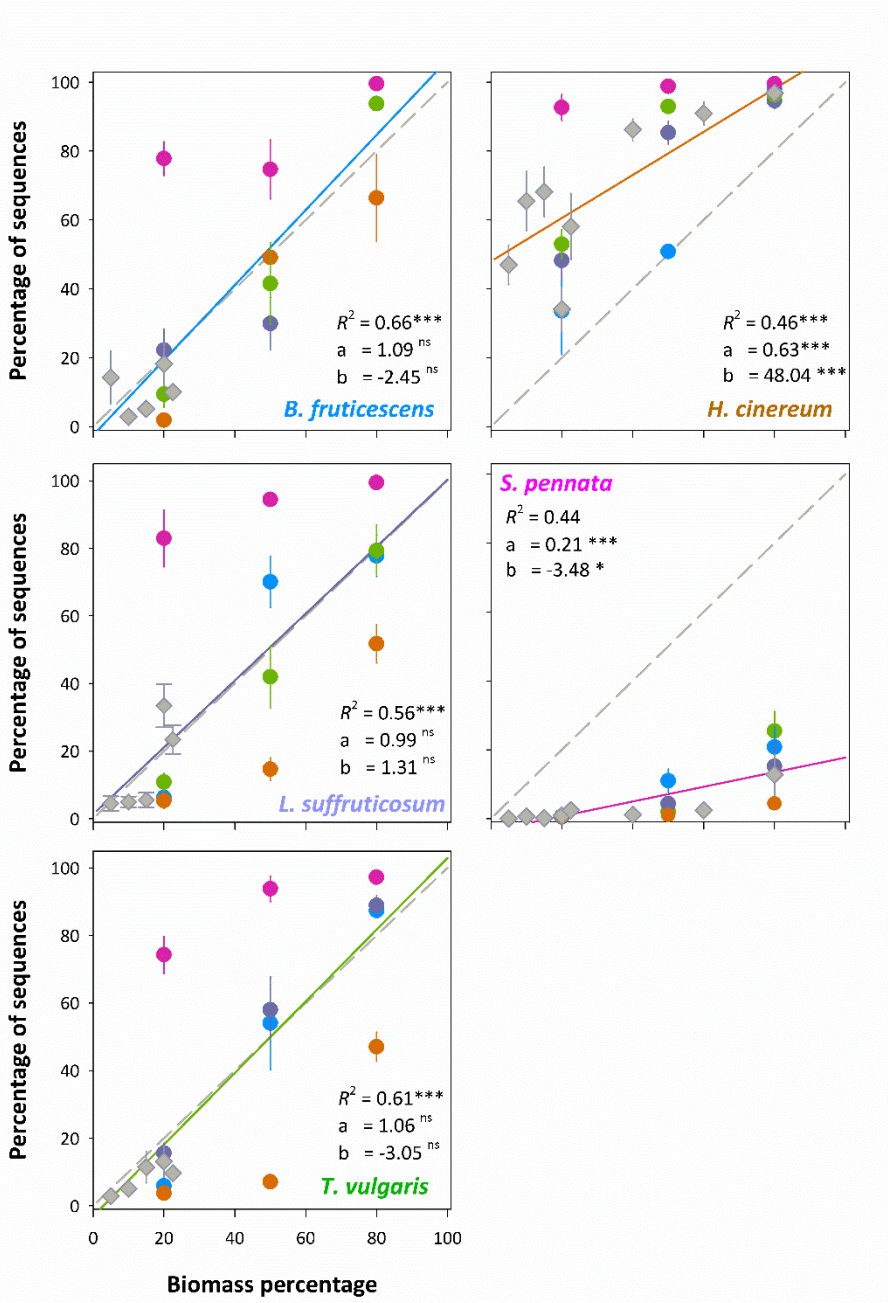


Figure 3. Correction indices for each species calculated from the pairwise mixtures, based on the number of reads retrieved after sequencing and the percentage of biomass proportion in each mixture. For each species, a correction index was calculated using the mixtures where the species was at 20, 50 and 80% of biomass. Avg. refers to the average correction index. The dashed grey line in the top panels represents a correction index = 1. Note that the Y axes for the species in the top panels is the same.

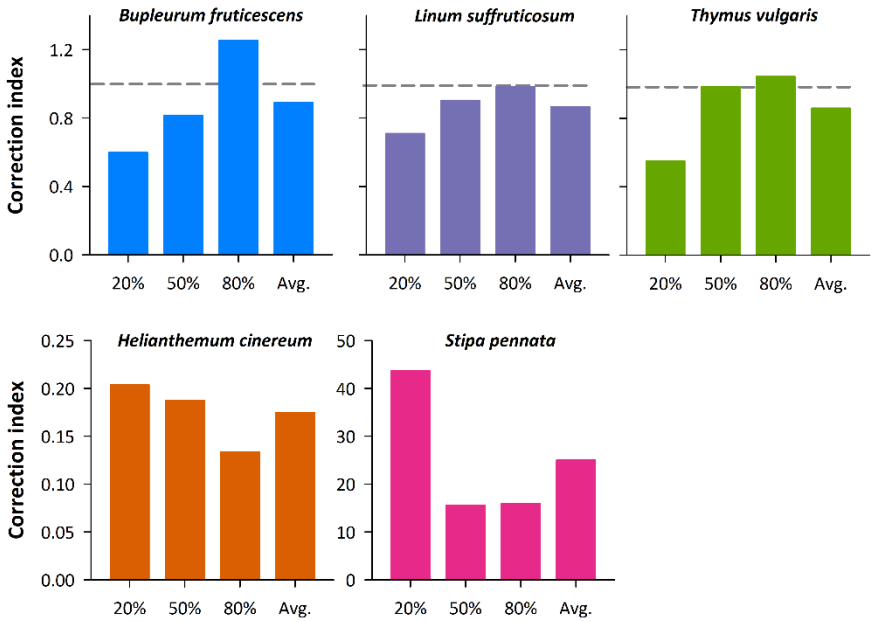


Figure 4. Comparison of the percentage of biomass of each species in the multispecies mixtures (left bars) to the observed (uncorrected; central bars) and the corrected (after recalculation of the number of reads with the 50% correction indices; right bars) percentage of reads. Panels a) to i) show a specific type of multispecies mixture (defined by the left column), and each color represents the percentage of biomass/reads proportion of each species, averaged for the three replicate samples of each mixture. Gray stacks in the uncorrected bars represent the proportion of sequences retrieved from species other than those included in the mixtures.

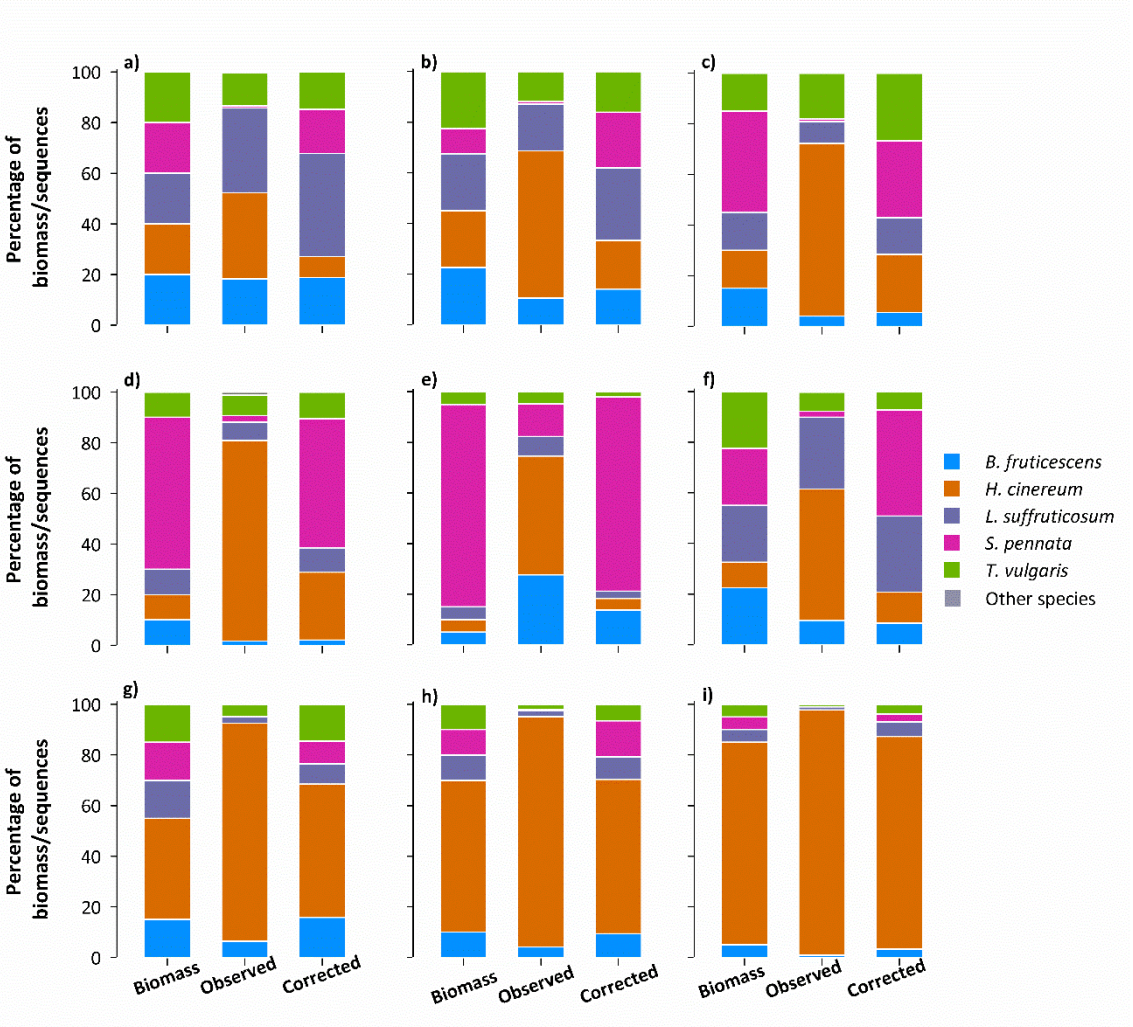
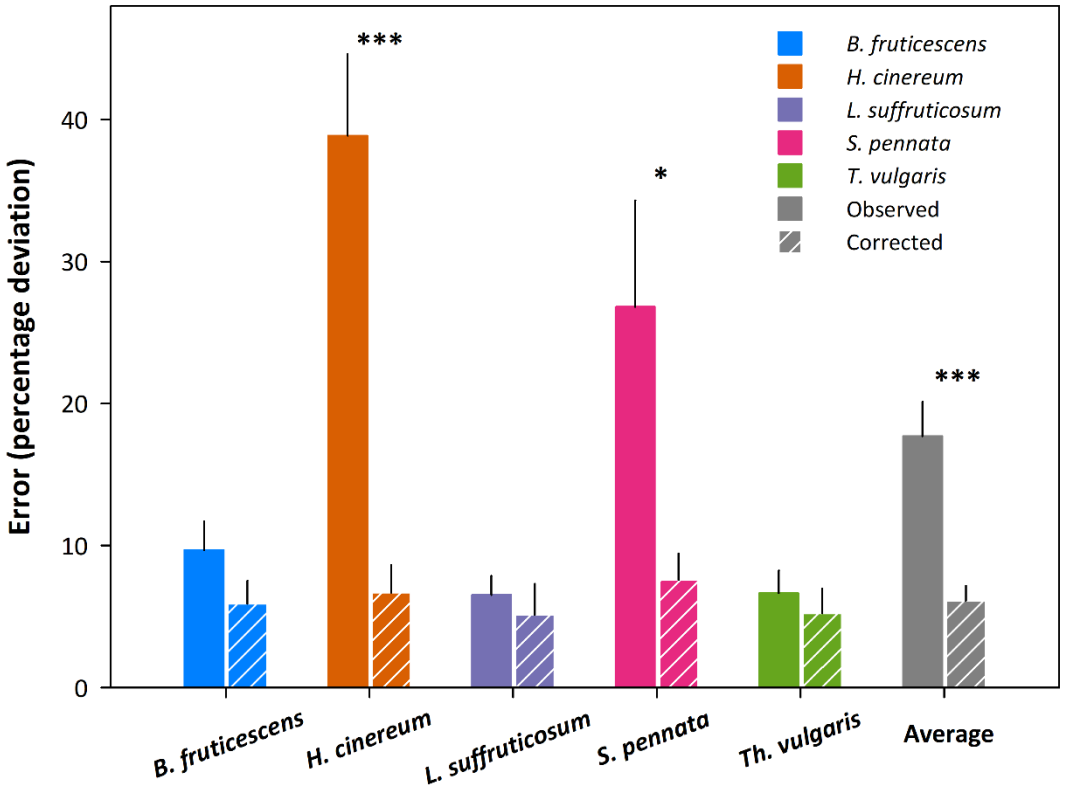


Figure 5. Average error (average absolute difference between the real percentage of biomass and the corrected/uncorrected percentages of reads) for each species. Filled bars: error for uncorrected reads. Striped bars: error for corrected reads. Lines represent 1 s.e. Significant differences between corrected and uncorrected deviations (one-way ANOVA) are indicated by asterisks (*** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$).



MOLECULAR ECOLOGY RESOURCES

Supplemental Information for:

Estimating belowground plant abundance with DNA metabarcoding

Silvia Matesanz, David S. Pescador, Beatriz Pías, Ana M. Sánchez, Julia Chacón-Labella, Angela Illuminati, Marcelino de la Cruz, Jesús López-Angulo, Adrián Escudero

Table of Contents:

**

Instructions for Authors:

1. use of this branded Supplemental Information template is recommended, but not mandatory
2. consolidate your Supplemental files into as few documents as possible
3. if your file of Supplemental Information is very large, create a Table of Contents
4. Your Supplemental Information will not be copy-edited. Do not leave in track-changes and other editing marks. The document will be posted "as is."
5. Save as PDF if possible

MOLECULAR ECOLOGY
RESOURCES

Table S1. Species composition and biomass proportion of each mock community (left). The right section of the table shows the positive identification of the species after sequencing (in green), and the detection of species during sequencing that were not added in the community (in red). Comm.: Community; B_s: *Bupleurum fruticescens*; H_c: *Helianthemum cinereum*; L_s: *Linum suffruticosum*; S_p: *Stipa pennata*; T_v: *Thymus vulgaris*; S_l: *Salvia lavandulifolia*; C_m: *Coris monspeliensis*; Q_sp: *Quercus sp*; T_d: *Thesium divaricatum*; L_c: *Leuzea conifera*.

Sample	B_s	H_c	L_s	S_p	T_v		B_s	H_c	L_s	S_p	T_v	S_l	C_m	Q_sp	T_d	L_c
A_01	50%		50%				38.50		61.50							
A_02	50%		50%				14.80		85.20							
A_03	50%		50%				36.50		63.50							
A_04	50%				50%		17.70		1.00		81.30					
A_05	50%				50%		53.22	1.81	1.31		37.61				6.05	
A_06	50%				50%		53.80		3.00		43.30					
A_07	50%	50%					45.40	54.60								
A_08	50%	50%					49.80	50.20								
A_09	50%	50%					52.20	47.80								
A_10		50%	50%					89.01	10.99							
A_11		50%	50%					78.63	21.37							
A_12		50%	50%					88.38	11.62							
A_13		50%			50%			94.54			5.46					
A_14		50%			50%			92.57			7.43					
A_15		50%			50%			91.65			8.35					
A_16	50%			50%			83.41	0.23	10.34	5.41	0.28					0.33
A_17	50%			50%			83.05		0.45	10.44					3.58	2.49
A_18	50%			50%			57.55		19.48	17.12	0.22					5.62
A_19		50%		50%				99.00		0.57	0.43					
A_20		50%		50%				98.55		1.45						
A_21		50%		50%				100.00		0.00						
A_22			50%	50%						95.54	3.99	0.47				
A_23			50%	50%						94.56	3.50	1.94				
A_24			50%	50%						93.25	5.62	1.12				
A_25				50%	50%			0.86		3.37	95.77					
A_26				50%	50%					0.87	99.13					
A_27				50%	50%			11.73		1.77	86.49					
A_28			50%		50%				25.05		74.95					
A_29			50%		50%				56.79		43.21					
A_30			50%		50%				44.12		55.88					
B_01	20%	20%	20%	20%	20%		24.47	26.03	37.22	1.09	10.90	0.29				
B_02	20%	20%	20%	20%	20%		6.08	29.47	42.08	0.51	21.85					

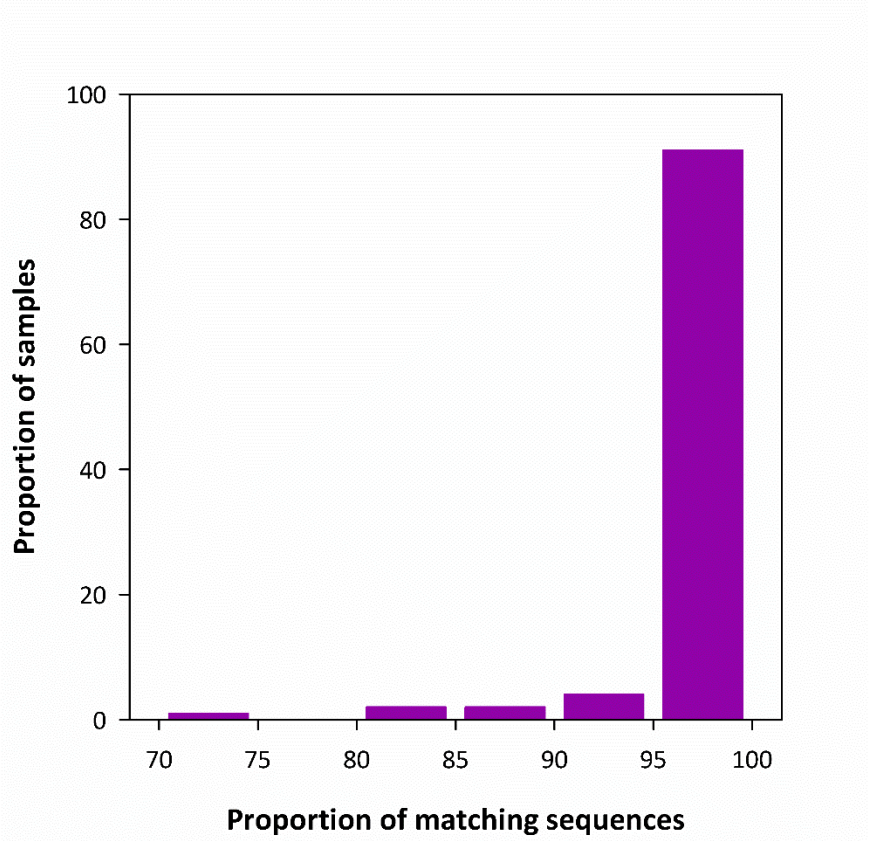
B_03	20%	20%	20%	20%	20%	24.06	46.81	20.95	1.07	6.53	0.57
C_01	22.50%	22.50%	22.50%	10%	22.50%	8.50	71.34	13.49	0.37	6.31	
C_02	22.50%	22.50%	22.50%	10%	22.50%	12.01	39.26	27.45	1.21	20.07	
C_03	22.50%	22.50%	22.50%	10%	22.50%	11.25	63.55	14.63	1.22	9.34	
C_04	15%	15%	15%	40%	15%	5.36	67.53	15.80	1.30	10.01	
C_05	15%	15%	15%	40%	15%	2.06	80.96	5.48	0.94	10.56	
C_06	15%	15%	15%	40%	15%	4.59	56.00	4.69	1.16	33.56	
C_07	10%	10%	10%	60%	10%	1.11	80.74	10.42	3.45	4.27	
C_08	10%	10%	10%	60%	10%	2.96	64.82	9.47	3.24	16.01	3.50
C_09	10%	10%	10%	60%	10%	0.94	92.01	2.18	0.74	4.13	
C_10	5%	5%	5%	80%	5%	8.36	45.86	15.03	24.78	5.98	
C_11	5%	5%	5%	80%	5%	45.49	37.77	4.31	6.64	5.79	
C_12	5%	5%	5%	80%	5%	29.00	57.31	3.93	7.03	2.74	
C_13	22.50%	10%	22.50%	22.50%	22.50%	5.80	73.96	15.00	1.70	3.54	
C_14	22.50%	10%	22.50%	22.50%	22.50%	11.34	42.99	35.24	3.34	6.93	0.16
C_15	22.50%	10%	22.50%	22.50%	22.50%	11.63	38.64	34.84	2.33	11.87	0.70
C_16	40%	15%	15%	15%	15%	16.22	80.97	1.19	0.17	1.45	
C_17	40%	15%	15%	15%	15%	1.92	85.79	4.64	0.00	7.65	
C_18	40%	15%	15%	15%	15%	1.01	91.79	1.50	0.37	5.33	
C_19	60%	10%	10%	10%	10%	0.66	95.64	1.97	0.00	1.73	
C_20	60%	10%	10%	10%	10%	11.02	84.38	1.79	0.63	2.18	
C_21	60%	10%	10%	10%	10%	0.78	92.76	3.84	0.33	2.30	
C_22	80%	5%	5%	5%	5%	0.95	95.81	1.58	0.00	1.66	
C_23	80%	5%	5%	5%	5%	1.58	96.05	1.45	0.15	0.77	
C_24	80%	5%	5%	5%	5%	0.19	98.69	0.73	0.00	0.39	
D_01	20%		80%			31.85		68.15			
D_02	20%		80%			24.08		75.92			
D_03	20%		80%			10.81		89.19			
D_04	20%				80%	15.67				84.33	
D_05	20%				80%	2.52		7.76		89.72	
D_06	20%				80%	10.09	0.25	1.65		88.00	
D_07	20%	80%				43.58	56.42				
D_08	20%	80%				87.10	12.90				
D_09	20%	80%				68.52	31.48				
D_10		20%	80%				36.90	63.10			
D_11		20%	80%				54.20	45.80			
D_12		20%	80%				53.61	46.39			
D_13		20%			80%		61.34			38.66	
D_14		20%			80%		46.93			53.07	
D_15		20%			80%		50.61			49.39	

D_16	20%			80%		0.13			84.80				15.06
D_17	20%			80%		99.48			0.52				
D_18	20%			80%		99.60			0.40				
D_19		80%		20%			100.00		0.00				
D_20		80%		20%			100.00		0.00				
D_21		80%		20%			99.52		0.48				
D_22			80%	20%				99.22	0.78				
D_23			80%	20%				99.70	0.30				
D_24			80%	20%				99.38	0.62				
D_25				20%	80%				0.99	99.01			
D_26				20%	80%		6.24		0.28	93.48			
D_27				20%	80%				0.67	99.20	0.13		
D_28			80%		20%			89.89		10.11			
D_29			80%		20%			83.55		16.45			
D_30			80%		20%		15.27	64.53		20.20			
D_31	80%		20%			95.52		4.48					
D_32	80%		20%			90.68		9.32					
D_33	80%		20%			94.97		5.03					
D_34	80%				20%	94.55		0.96		4.49			
D_35	80%				20%	90.30				9.70			
D_36	80%				20%	96.52				3.48			
D_37	20%	80%				0.89	99.11						
D_38	20%	80%				0.70	99.30						
D_39	20%	80%				4.08	95.92						
D_40		80%	20%				97.60	2.40					
D_41		80%	20%				90.19	9.81					
D_42		80%	20%				96.11	3.89					
D_43		80%			20%		94.57			5.43			
D_44		80%			20%		96.56			3.44			
D_45		80%			20%		97.62			2.38			
D_46	20%			80%		87.56	0.17		8.85	3.42			
D_47	20%			80%		73.63		0.41	25.74	0.22			
D_48	20%			80%		72.32			27.68				
D_49		20%		80%			84.98		6.74	7.74	0.54		
D_50		20%		80%			96.30		3.70				
D_51		20%		80%			96.66		2.73	0.61			
D_52			20%	80%			0.11	90.35	8.62	0.92			
D_53			20%	80%				66.11	29.72	2.42			1.75
D_54			20%	80%			0.33	92.41	7.25				
D_55				80%	20%				36.12	63.88			

D_56		80%	20%				0.94	16.62	82.44
D_57		80%	20%					23.54	76.46
D_58	20%		80%			0.94	15.81		83.26
D_59	20%		80%				9.79		90.21
D_60	20%		80%				6.85		93.15

MOLECULAR ECOLOGY
RESOURCES

Figure S1. Proportion of samples according to the proportion of sequences assigned to the species added in each mock community.



MOLECULAR ECOLOGY
RESOURCES

Table S2. Intercept, slope, p-values for t-tests (hypotheses: intercept = 0 and slope = 1) and adjusted R^2 of the fit between the percentage of biomass in the samples and the percentage of reads (DNA sequences).

Pairwise samples	Intercept	*P*-value Intercept test	Slope	*P*-value Slope test	Adj. R^2
Bupleurum fruticescens	5.224	0.576	0.9873	0.940	0.501
Helianthemum cinereum	44.6472	**<0.00001**	0.666924	**0.020**	0.411
Linum suffruticosum	10.6417	0.346	0.8453	0.444	0.325
Stipa pennata	-6.32159	**0.039**	0.26837	**<0.00001**	0.427
Thymus vulgaris	6.6994	0.543	0.9209	0.688	0.377
All samples					
Bupleurum fruticescens	-2.44947	0.566	1.08932	0.941	0.660
Helianthemum cinereum	48.1151	**<0.00001**	0.6336	**0.012**	0.470
Linum suffruticosum	1.3135	0.786	0.9898	0.442	0.560
Stipa pennata	-3.59539	**0.018**	0.21154	**<0.00001**	0.442
Thymus vulgaris	-3.0526	0.518	1.0597	0.687	0.603

MOLECULAR ECOLOGY
RESOURCES

Table S3. Correction factors for each species using one control species (three first columns, following Thomas et al. 2016 Mol. Ecol. Resources 16), and using all species mixtures (last column).

Species	Correction index using *Linum* as control species	Correction index using *Stipa* as control species	Correction index using *Helianthemum* as control species	Correction index based on all pairwise mixtures
Bupleurum fruticescens	2.4	0.192	1.026	0.817
Helianthemum cinereum	0.056	0.011	--	0.188
Stipa pennata	3.22	--	88.053	15.506
Linum suffruticosum	--	0.047	5.845	0.904
Thymus vulgaris	0.728	0.021	13.116	0.985

MOLECULAR ECOLOGY RESOURCES

Table S4. Comparison of the percentage of biomass of each species in the multispecies mixtures (actual biomass) to the observed (uncorrected) and the corrected (after recalculation of the number of reads with correction indices) percentage of reads. Correction indices were computed using correction indices calculated with one control species (Linum, Stipa or Helianthemum, see Table S3), or using all species. The average error is calculated for the three replicates of each multispecies mixture as the absolute difference between the corrected and uncorrected biomass percentages for each species.

Species	***B. fruticescens***	***H. cinereum***	***L. suffruticosum***	***S. pennata***	***Th. vulgaris***	**Average error**
Mixture 1. Actual Biomass (%)	5.00	5.00	5.00	80.00	5.00	
Uncorrected biomass (%)	27.62	46.98	7.76	12.82	4.84	26.94
Corrected biomass (*Linum* as control sp.)	54.15	2.27	6.42	34.32	2.84	20.23
Corrected biomass (*Stipa* as control sp.)	33.75	3.12	1.71	60.86	0.56	11.50
Corrected biomass (*Helianthemum* as control sp.)	3.44	4.77	3.31	82.55	5.94	1.40
Corrected biomass (using all spp.)	13.72	4.58	2.94	76.66	2.09	3.49
Mixture 2. Actual Biomass (%)	**22.50**	**22.50**	**22.50**	**10.00**	**22.50**	
Uncorrected biomass (%)	10.59	58.05	18.53	0.93	11.91	14.22
Corrected biomass (*Linum* as control sp.)	44.14	6.16	30.90	4.92	13.88	12.02
Corrected biomass (*Stipa* as control sp.)	43.23	15.00	17.98	18.67	5.12	11.76
Corrected biomass (*Helianthemum* as control sp.)	2.75	16.36	26.09	19.19	35.61	10.35
Corrected biomass (using all spp.)	14.05	19.27	28.60	21.93	16.15	7.21
Mixture 3. Actual Biomass (%)	**20.00**	**20.00**	**20.00**	**20.00**	**20.00**	
Uncorrected biomass (%)	18.21	34.11	33.42	0.89	13.09	11.07
Corrected biomass (*Linum* as control sp.)	45.90	2.16	37.67	3.08	11.18	17.43
Corrected biomass (*Stipa* as control sp.)	49.19	6.14	26.36	13.21	5.10	14.22
Corrected biomass (*Helianthemum* as control sp.)	4.19	7.48	38.56	17.17	32.59	12.46
Corrected biomass (using all spp.)	18.78	8.25	40.80	17.45	14.73	8.32
Mixture 4. Actual Biomass (%)	15.00	15.00	15.00	40.00	15.00	
Uncorrected biomass (%)	4.00	68.16	8.66	1.13	18.04	22.48
Corrected biomass (*Linum* as control sp.)	23.98	11.04	22.41	9.74	32.82	13.69

Corrected biomass (*Stipa* as control sp.)	21.43	23.38	11.22	32.84	11.14	5.92
Corrected biomass (*Helianthemum* as control sp.)	0.90	16.46	12.11	22.88	47.65	13.64
Corrected biomass (using all spp.)	5.43	22.96	14.56	30.26	26.79	7.90

Mixture 5. Actual Biomass (%)	10.00	10.00	10.00	60.00	10.00	
Uncorrected biomass (%)	1.67	79.19	7.36	2.48	8.14	27.91
Corrected biomass (*Linum* as control sp.)	13.40	19.16	23.21	25.13	19.10	13.95
Corrected biomass (*Stipa* as control sp.)	7.73	26.70	7.49	53.99	4.10	6.68
Corrected biomass (*Helianthemum* as control sp.)	0.38	22.49	8.91	45.17	23.05	10.21
Corrected biomass (using all spp.)	2.00	26.88	9.62	51.01	10.50	6.95

Mixture 6. Actual Biomass (%)	22.50	10.00	22.50	22.50	22.50	
Uncorrected biomass (%)	9.59	51.86	28.36	2.46	7.45	19.15
Corrected biomass (*Linum* as control sp.)	34.00	5.27	41.04	11.97	7.72	12.01
Corrected biomass (*Stipa* as control sp.)	28.50	10.52	20.19	38.35	2.43	8.95
Corrected biomass (*Helianthemum* as control sp.)	1.79	11.09	29.71	40.04	17.38	10.33
Corrected biomass (using all spp.)	8.55	12.25	30.02	41.95	7.24	11.69

Mixture 7. Actual Biomass (%)	15.00	40.00	15.00	15.00	15.00	
Uncorrected biomass (%)	6.38	86.18	2.44	0.18	4.81	18.47
Corrected biomass (*Linum* as control sp.)	41.56	23.66	12.26	3.20	19.33	12.35
Corrected biomass (*Stipa* as control sp.)	34.95	45.37	5.93	8.22	5.53	10.13
Corrected biomass (*Helianthemum* as control sp.)	4.48	47.75	7.31	8.93	31.52	9.71
Corrected biomass (using all spp.)	15.87	52.56	8.03	8.99	14.55	5.37

Mixture 8. Actual Biomass (%)	10.00	60.00	10.00	10.00	10.00	
Uncorrected biomass (%)	4.15	90.93	2.53	0.32	2.07	12.37
Corrected biomass (*Linum* as control sp.)	33.85	34.69	17.44	4.40	9.63	12.51
Corrected biomass (*Stipa* as control sp.)	24.42	55.27	6.42	11.63	2.26	6.42
Corrected biomass (*Helianthemum* as control sp.)	2.31	56.92	9.06	15.16	16.55	4.68
Corrected biomass (using all spp.)	9.39	61.01	8.84	14.28	6.47	2.12

Mixture 9. Actual Biomass (%)	5.00	80.00	5.00	5.00	5.00	
Uncorrected biomass (%)	0.91	96.85	1.26	0.05	0.94	6.74
Corrected biomass (*Linum* as control sp.)	20.27	58.97	12.65	1.34	6.78	9.88
Corrected biomass (*Stipa* as control sp.)	11.67	79.79	4.13	2.99	1.43	2.67
Corrected biomass (*Helianthemum* as control sp.)	0.73	80.19	5.92	3.31	9.85	2.38
Corrected biomass (using all spp.)	3.27	84.08	5.61	3.17	3.86	1.88

MOLECULAR ECOLOGY
RESOURCES

Figure S2. Average error (average absolute difference between the actual percentage of biomass and the corrected/uncorrected percentages of reads) for each multispecies mixture (averaged across species). Corrections were performed based on different correction indices (see Table S3). Species-specific corrected biomass can be found on Table S4. The composition and relative abundance of each multispecies mixture (code numbers in the X axis match those on Table S4).