msGBS: A new high-throughput approach to quantify the relative species abundance in root samples of multi-species plant communities

Authors

C.A.M. Wagemaker[1], L. Mommer[2], E.J.W. Visser[1], A. Weigelt[3,4], T.P. van Gurp[5], M. Postuma[2], A.E. Smit-Tiekstra[1], H. de Kroon[1]

Publon - Web of Science ResearcherID : N-6570-2018

Affiliations:

[1] Institute for Water and Wetland Research, Department of Experimental Plant Ecology, Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
[2] Plant Ecology and Nature Conservation Group, Wageningen University and Research, PO box 47, 6700 AA, Wageningen, The Netherlands
[3] Systematic Botany and Functional Biodiversity, Faculty of Life Sciences, University of Leipzig, Leipzig, Germany
[4] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany
[5] Naktuinbouw, Sotaweg 22, 2371 GD Roelofarendsveen, The Netherlands

Keywords

DNA, High Throughput Sequencing, Genotyping By Sequencing, roots, biodiversity, species abundance

Corresponding author

C.A.M. Wagemaker, Tel: +31(0)243652919, E-mail: n.wagemaker@science.ru.nl

Abstract

Plant interactions are as important belowground as aboveground. Belowground plant interactions are however inherently difficult to quantify, as roots of different species are difficult to disentangle. Although for a couple of decades molecular techniques have been successfully applied to quantify root abundance, root identification and quantification in multi-species plant communities remains particularly challenging.

Here we present a novel methodology, multi-species Genotyping By Sequencing (msGBS), as a next step to tackle this challenge. First, a multi-species meta-reference database containing thousands of gDNA clusters per species is created from GBS derived High Throughput Sequencing (HTS) reads. Second, GBS derived HTS reads from multi-species root samples are mapped to this meta-reference which, after a filter procedure to increase the taxonomic resolution, allows the parallel quantification of multiple species.

The msGBS signal of 111 mock-mixture root samples, with up to 8 plant species per sample, was used to calculate the within-species abundance. Optional subsequent calibration yielded the across-species abundance. The within- and across-species abundances highly correlated (R^2 range 0.72-0.94 and 0.85-0.98, respectively) to the biomass-based species abundance. Compared to a qPCR based method which was previously used to analyze the same set of samples, msGBS provided similar results. Additional data on 11 congener species groups within 105 natural field root samples showed high taxonomic resolution of the method.

msGBS is highly scalable in terms of sensitivity and species numbers within samples, which is a major advantage compared to the qPCR method and advances our tools to reveal hidden belowground interactions.

Introduction

Our understanding of root distributions is limited compared to our knowledge of the patterning of leaves and shoots. This difference is largely due to methodological challenges as roots of different species can generally not be identified visually. With the introduction of DNA-based detection techniques (e.g. Bobowski et al., 1999; Jackson et al., 1999; Linder et al., 2000; Mommer et al., 2011; Jones et al, 2011), the first steps were taken in opening the 'black box of the underground'. Until 2008 these techniques were based on classic PCR amplification of nuclear, chloroplast or mitochondrial plant barcode loci, often combined with Sanger sequencing or RFLP (e.g. Bobowski et al., 1999; Jackson et al., 1999; Brunner et al., 2001; Ridgway et al., 2003; Wildová et al., 2004; McNickle et al., 2008). Individual root segments were identified on the basis of obtained PCR product length, DNA sequence or RFLP pattern. In some studies the species abundances were estimated after identification of numerous single root segments isolated from a single root core (Frank et al., 2015; Kesanakurti et al., 2011).

Mommer et al. (2008) and McKay et al. (2008) were the first to introduce Quantitative Polymerase Chain Reaction (qPCR) in studies on plant root distributions. Rather than extracting DNA from individual root segments, Mommer et al. (2008) extracted DNA from multispecies root samples. In a 4-species model system, the across-species abundance of root samples was estimated by relating the qPCR signals from root mixtures of unknown assembly to the qPCR signals of hand-mixed root samples of equal biomass proportions (i.e. calibration samples). In addition, species-specific primers, rather than universal primers were used. This method was later successfully applied in biodiversity experiments using plant mixtures with up to 8 species (e.g. Hendriks et al., 2015; Mommer et al., 2010; Oram et al., 2018; Zeng et al., 2017).

Although many successful uses, there are three main drawbacks of using qPCR all connected to the use of species-specific primers; 1) the primer development for each new species and the increased difficulty of it if species are more related 2) the variable sensitivity of these primers and 3) each species has to be analyzed separately. These drawbacks inspired us to explore the use of High Throughput Sequencing (HTS) for the quantification of relative species abundance in mixed root samples.

DNA sequence identification and counts can be used for both species- identification and quantification. Hiiesalu and colleagues (2012) were first to apply HTS in the field of root ecology, using the 454 Life Sciences sequencing platform. Hiiesalu et al. (2012) showed the power of HTS, but the use of a single barcoding marker resulted in insufficient taxonomic

91   resolution; the 37 species identified aboveground were represented by 29 belowground
92   Molecular Operational Taxonomic Units (MOTUs). Matesanz (2019) sequenced a 517bp
93   chloroplast rbcl marker using Miseq to analyze the root proportions of five shrubland dominant
94   species but recorded insufficient biomass versus sequence reads correlations and high false
95   positive rates. Lang et al. (2019) used the combined sequence information of 65 to 71
96   chloroplast Protein Coding Genes (PCG) within 'genome skims' (low-coverage, short-read
97   sequence datasets) to estimate pollen donor proportions within pollen mixtures using. However,
98   for two out of six pollen donor species the taxonomic resolution was still insufficient. The use
99   of genome skims to map to a small set of genes is very data inefficient, even more when applied
100 on roots which contain much lower number of plastids (Bramham & Pyke, 2017). Peel et al.
101 (2019) described RevMet; 49 wild reference species were represented by 'genome skims'
102 which were mapped to individual long Minion sequence reads derived from mixed species
103 pollen samples. Each read was assigned to a plant species and species proportions calculated
104 from the collection of identified reads. The method was validated using 6 replicate mock pollen
105 mixtures of known composition. This elegant approach shows promise but struggled with false
106 positive assignments within one of the two congener plant species pairs. Root and bee pollen
107 grains have in common that they host many Fungi (Brundrett, 2004; Leidenfrost et al., 2020)
108 which influence the taxonomic resolution and the quantification of plant species proportions.
109 Ondov et al. (2019) introduced Mash Screen, a MinHash (Ondov et al., 2016) based approach
110 which enables containment estimates for every NCBI RefSeq genome within every SRA
111 metagenome. Mash Screen has not been validated for quantification of species abundances in
112 plant mixtures but has great potential. While current tests are still limited, the results so far
113 suggest that none of the currently available methods is able to accurately identify all species in
114 mixed samples.

115   In this paper we describe the application of Genotyping By Sequencing (GBS)(Elshire
116 et al., 2011) on multiple species root samples (msGBS) which, combined with a gDNA cluster
117 filtering strategy, has the potential of increasing taxonomic resolution. GBS is developed for
118 SNP detection; gDNA is fragmented using endonucleases and a set of two synthetic dsDNA
119 adapters ligated to the fragments. Due to this preparation only a subset of the full genome is
120 PCR amplified. GBS provides a middle ground between targeted- and whole-genome shotgun
121 barcoding. The sequenced subset is clustered into a relative small reference genome which, in
122 msGBS, is enriched for species unique clusters increasing the taxonomic resolution.

The msGBS method we developed was aimed for two purposes:

1) quantify within-species abundance in mixed root samples in one single molecular analysis with unprecedented taxonomic resolution;

2) link the within-species abundance to root biomass across-species abundance using the calibration procedure *sensu* Mommer et al. (2008).

Material and methods

Experimental setup – Jena field study

The root samples used in this study were derived from the Jena Trait-Based Experiment (Barry et al., 2019; Ebeling et al., 2014; Oram et al., 2018), with two separate species pools. Pool 1 consisted of four forbs (*Centaurea jacea* L.*, Knautia arvensis* (L.) Coult.*, Leucanthemum vulgare* Lam.*, Plantago lanceolata* L.) and four grasses (*Festuca rubra* L.*, Helictotrichon pubescens* Huds.*, Phleum pratense* L.*, Poa pratensis* L.). Pool 2 also consisted of four forbs (*Geranium pretense* L.*, Leucanthemum vulgare, Plantago lanceolata, Ranunculus acris* L.) and four grasses (*Anthoxanthum odoratum* L.*, Dactylis glomerata* L.*, Holcus lanatus* L.*, Phleum pratense*). Three species, the forbs *Leucanthemum* and *Plantago* and the grass *Phleum*, were present in both pools (Figure 1). Monoculture and mixed field plots from both pools with up to 8 species were originally studied. All plots were mown twice yearly and weeded three times a year. Root cores of both pools were collected in 2016 (Oram et al., 2018) and carefully washed (debris, seeds, tubers, stolon's and taproots were carefully removed). The monoculture root material was used for the assembly of 13 'monoculture', 20 'calibration' and 111 'mock-mixture' samples:

- The ten calibration samples per species pool were assembled from monoculture root material in equal per species proportions.
- 56 and 55 mock-mixture samples of pool 1 and 2, respectively, were assembled from monoculture root material of each of eight species per pool and varied in proportions from 0 to 50 percent.

gDNA was extracted for these and the 'unknown' mixed field plot samples and subsequently analyzed by qPCR to quantify relative fine root abundances according to Mommer et al. (2008). Only the 'monoculture', 'calibration' and 'mock-mixture' samples were processed using msGBS (Figure 1). The monoculture samples were processed to assemble the meta-reference and used for downstream meta-reference filtering, the calibration samples were used to calibrate the within-species abundance to across-species abundance and the mock-mixture samples were used for the evaluation of msGBS in terms of correlations (to weighed root biomass and qPCR) and False- Positive and Negative Signals (FPS and FNS). Based on the FPS an analytical detection limit can be introduced (Alberdi et al., 2018; Garrido-Sanz et al., 2020). This part is further referred to as the '*Jena field study*'.

162   *Experimental setup – Dutch field study*

163   For the evaluation of the taxonomic resolution of msGBS we analyzed the msGBS

164   relative FPS (rFPS) of 11 congener groups within a field experiment, further referred to as the

165   '*Dutch field study*'. In this field study aboveground vegetation surveys were compared to the

166   belowground non-calibrated msGBS within-species abundances. Leaves of in total 120 plant

167   species (Table S1) were collected from 7 field sites across a 30 km trajectory along the main

168   branch of the river Rhine dike grasslands between the villages Ooij and Tiel in The Netherlands

169   for meta-reference creation. A Braun-Blanquet (Braun-blanquet, 1932) vegetation survey was

170   performed at two levels in each of the 7 field sites. A 5x5m2 plot survey and 5 1x1m2 plots

171   within the broader plot. From each of the 1x1m2 plots two 40x400mm root cores were taken

172   and were subdivided in 0-10cm, 10-20cm and 20-40cm depth portions, the replicate samples

173   were combined which, after careful root washing, totaled to 105 'field mixture root' samples.

174   gDNA was extracted from all leaf and root samples. The collected survey data and non-

175   calibrated msGBS results was used to assess the congener (Table S1) msGBS relative False

176   Positive Signals (rFPS).

177

178   *gDNA extractions and qPCR*

179   gDNA of the *Jena field study* samples were previously extracted using the DNeasy plant

180   kit (Qiagen, The Netherlands). The qPCR methodology and root distributions were previously

181   described in the Jena Trait-Based Experiment papers (Barry et al., 2019; Oram et al., 2018).

182   gDNA of the *Dutch field study* samples was extracted using the Nucleospin® plant II kit (MN,

183   Germany).

184

185   *msGBS library preparations and sequencing*

186   The GBS protocol, as described by Elshire et al. (2011), was altered regarding the restriction

187   enzymes and the adapter design (Fig. S1 and Table S2). A more detailed lab protocol can be

188   found in the Extended lab protocol section of the supporting information. In total 3 pooled

189   sequence libraries were constructed; one for the 144 samples of the *Jena field study*, one for the

190   122 monoculture leaf samples of the *Dutch field study* and one for the 105 root samples of the

191   *Dutch field study*. The *Dutch field study* samples were equimolar pooled using qPCR. Half a

192   sequence run (lane) was used for the Jena msGBS library and a full sequence run for each of

193   the *Dutch field study* msGBS libraries.

194   First, 300 ng of genomic DNA (gDNA) of each sample was digested by two restriction

195   enzymes (*Pac*I and *Nsi*I) after which two indexed adapters were ligated to the DNA fragments.

196   The main change in the adapter design was the incorporation of 3 random nucleotides per

197   adapter for the identification of PCR duplicates within each amplified msGBS library. After the

198   ligation step the samples were pooled, mixed and aliquoted in eight portions per library for

199   practical reasons and to prevent the effect of PCR bias. For the *Dutch field study* the diglig

200   reactions were equimolar pooled based on a qPCR quantification using the KAPA Library

201   Quantification Kit for HTS (KAPA Biosystems, USA). The aliquots were concentrated

202   (QIAquick, Qiagen), AMPureXP size selected preferring >150bp DNA fragments (Beckman

203   coulter, Canada) and PCR amplified using KAPA HiFi HotStart readyMix (Roche Diagnostics,

204   Switzerland). The PCR reactions were combined, QIAquick concentrated and quantified using

205   the KAPA Library Quantification Kit for HTS. The final three pooled msGBS libraries were

206   spiked with 10% PhiX DNA to increase the DNA complexity of the library in order to improve

207   the Hiseq color matrix estimation for which the first 11 sequencing cycles are used overlapping

208   with our index region. Sequencing was performed by Novogene (Hongkong) on a Illumina

209   (USA) Hiseq X-Ten sequencer; 2x150bp Paired-End (PE) sequencing reads, each one starting

210   with 3 random nucleotides and the adapter index.

211

212   *msGBS data processing*

213   Computations were executed on a local Linux cluster node in Nijmegen, The

214   Netherlands. Writing and debugging of Python scripts (Python core Team, 2015) was

215   performed using PyCharm Professional 2017 2.2. R (Suhl et al., 2014) was executed using

216   Rstudio (RStudio Team, 2016). The msGBS data processing can be described by five processes

217   which are outlined in Figure 2;

218   Process 1: Sequence read pre-processing. First, the reads were demultiplexing; the

219   sequence read adapter indices are coupled to the sample name which was added to the read

220   header. The 2x3bp Unique Molecule Identifiers (UMI) nucleotides were processed and together

221   with the indices stripped from the sequence read and added to the read header (Fig. S1). Next,

222   the reads were inspected for adapter traces and low-quality nucleotides (<Q10) and trimmed

223   when needed. All PE reads were merged (minimum 20 bp overlap) or else joined. The combined

224   merged and joined reads are the assembled reads (Fig. 2, product 1).

225   Process 2: Meta-reference creation. For each monoculture a *de novo* assembled
226   reference was created from dereplicated and clustered (with 95% identity) monoculture
227   assembled reads. The clusters of all monoculture references are combined into a single meta-
228   reference (a digital gDNA sequence database) while retaining original monoculture identifier
229   names. The meta-reference was cleansed from all identifiable non-Eukaryota and Fungi clusters
230   by a local BLASTN search against the NCBI nr database (Fig. S3).

231   Process 3: Sequence read mapping. The assembled reads from *all* samples were mapped
232   to the meta-reference. A BAM (sequence alignment file) was created (Fig. 2, product 3), in
233   which the read header information was retained.

234   Process 4: Post-processing of read mapping data. First the UMI's and the mapping
235   alignment scores in the BAM file are processed; sequence reads are marked as 'is_duplicate'
236   or 'qc_fail', respectively. PCR duplicates are evaluated on a per meta-reference cluster, within
237   sample level. They can cause bias in the analysis as the duplication rate can vary between
238   amplified regions and samples. The BAM file is converted to CSV format; only the total read
239   counts per cluster per sample are retained, reads marked as 'is_duplicate' or 'qc_fail' are not
240   counted. A minimum total read count threshold of ten reads per cluster over *all* samples was
241   set; clusters that failed this criteria were removed from the CSV file.

242   An important step of the post-processing is the monoculture-based cluster filtering
243   which uses the monoculture read counts in the CSV file to identify and discard between
244   monoculture root sample homologous clusters. Removal of these clusters increases the
245   taxonomic resolution of msGBS. These homologous clusters are plant-born or non-plant-born
246   clusters that are present in multiple monoculture root samples. Monoculture per cluster read
247   counts were either 'target read counts' or 'non-target read counts' as illustrated in Figure 3.
248   The read counts are evaluated by three filter steps.

249   1) Prefilter; is the highest read count is indeed the target species.

250   2) Target count filter; a minimum of 8 counted reads (script filter parameter f1=8)
251   which enables effective non-target filtering (step 3).

252   3) Non-target count filter; the non-target read count threshold of the *Jena field study*
253   monocultures was set to 1/15th (script filter parameter f2=15) of the target read
254   count. This corresponds to a maximum non-target signal of 6.7%.

255   When a cluster passed all filter steps this cluster, and the reads counts of *all* samples, was
256   recorded in the filtered CSV file. Finally the total number of read counts, of all filtered clusters

257 combined, were counted for *all* samples. *Jena field study* samples for which, in total, less than
258 1000 reads were counted (script filter parameter f3=1000) were removed from the filtered CSV
259 file (Fig. 2, product 4).

260   Process 5: Non-calibrated and calibrated analysis. msGBS filtered CSV data was
261 processed in two steps as illustrated in Table 1. The first step, which was performed for both
262 the *Jena* and *Dutch field study* samples, is the non-calibrated analysis in which the per species
263 read counts is divided by the total reads count of the mock-mixture and field root mixture
264 samples, respectively. This resulted in the within-species abundance (Fig. 2, product 5a). The
265 second step, which was only performed for the *Jena field study*, is the optional calibration of
266 the within-species abundances. Since typical gDNA yields vary among species, we expected
267 biomass independent, species-specific variation in the number of reads within samples. To
268 estimate across-species abundance in mixed samples, the within-species abundance thus needed
269 to be calibrated (*sensu* Mommer et al. 2008). Ten calibration samples per pool, assembled from
270 per species equal proportions of fresh monoculture root biomass, were used to calculate a
271 calibration key. The calibration key was used to convert the within-species abundance of the
272 mock-mixture samples to across-species abundance (Fig. 2, product 5b) which was
273 subsequently projected on the total biomass.

274   *Jena field study* only; the FPS of the mock-mixture samples was evaluated by
275 calculating the averaged, per species, across-species abundance when this species was not
276 present in the assembly. The FNS threshold in calibrated mode was defined as 1% across-
277 species abundance. In non-calibrated mode the FNS threshold was defined on a per species
278 level; we defined this threshold as 1/50th of the msGBS signal of the, for that species, 50%
279 biomass mock-mixture samples.

280   *Dutch field study* only; we used the within-species abundances of 11 congener groups
281 to calculate the average rFPS, i.e., the average msGBS (field root mixture sample) signal of
282 absent species divided by the average msGBS signal of congener species that are present (Table
283 S3. The actual biomass-based proportions of these samples are unknown; samples were selected
284 for comparison when a) field plots were available in which not all species of a congener groups
285 were present and b) when a msGBS signal for the species that was present was detected.

286

287 More details on the bioinformatics can be found in the extended bioinformatics section of the
288 supporting information.

Statistics

290   Regression analysis were performed for all comparisons of the biomass-based species

291   proportions the qPCR- and msGBS estimates of relative species abundance. In order to evaluate

292   the between species variation in sequence read mapping counts of the calibration samples we

293   calculated the coefficient of variation (CV). A two-way ANOVA was used to test if the

294   calibrated msGBS and qPCR results were significantly different.

295

296   **Results**

297

298   *msGBS library preparations and sequencing*

299   For the *Jena field study* samples a total of 144 msGBS reactions were pooled into a

300   single msGBS sequencing library (Fig. S2) with a final DNA yield of 11.3 ng/µl (qPCR) and

301   an average fragment size of 940 basepairs (bp). The sequencing yielded 217,171,278 2x150bp

302   PE msGBS reads (Table S4). For the *Dutch field study* samples a total of 224 msGBS reactions

303   were equimolarly pooled in two subsets using qPCR quantification which resulted in two

304   msGBS sequencing libraries with a final yield of 15.7 ng/µ and 3.47 ng/µ, respectively. The

305   sequencing of these msGBS libraries resulted in 378,265,715 and 291,588,907 2x150bp PE

306   msGBS reads, respectively (Table S4).

307

308   *Jena field study msGBS results*

309   The results of the *Jena field study* msGBS data processing following Figure 2.

310   Process 1. Sequence pre-processing. During demultiplexing of the monoculture-,

311   calibration- and mock-mixture samples adapter barcodes were successfully identified in

312   181,555,188 reads (84%). This number ranged from 67,657 to 4,505,442 per sample. Non-

313   identified reads originated from PhiX DNA (10.9%) or adapter dimers (6.1%). Adapter traces

314   were identified and trimmed in 13.8% of the reads. Of the assembled reads (Fig. 2, product 1)

315   39% and 61% of the reads were merged and joined, respectively.

316   Process 2. *De novo* meta-reference creation. On average 102,155 meta-reference

317   clusters were generated per monoculture root sample; on average 1 cluster per 12 PE reads

318   which results in a total of 1,328,016 clusters (Table S5). The number of clusters per

319   monoculture varied from 8,475 to 260,885. A positive correlation (R^2=0.92) was found between

320   number of processed monoculture reads and generated clusters per species (Fig. S4) which

321   implies that a higher sequencing effort will result in more clusters per species. BLASTN

322 filtering removed 1.1% of the clusters from the meta-reference (Table S6) which were mainly
323 annotated as arbuscular mycorrhizal (AM) Fungi and Bacteria (60% and 34%, respectively).

324   Process 3. Sequence read mapping. In total 62% (Table S4) of the assembled reads of
325 monoculture-, calibration- and mock-mixture samples were recorded in the BAM file. That a
326 high proportion of reads did not map to the meta-reference can be caused by 1) the absence of
327 a homologous cluster due to low monoculture sequencing effort or BLASTN filtering or 2)
328 reads that were too short for mapping after low quality nucleotide trimming.

329   Process 4. On average 23.3% of the reads were marked as 'is-duplicate' and/or 'quality-
330 fail' in the BAM file. The extracted CVS file counted 86,443,367 reads (Table S4) which were
331 mapped to 726,605 clusters (Table S7).

332   The monoculture-based cluster filtering evaluated the read counts of the 9,446,804
333 monoculture reads (Table S4) to the 726,605 remaining clusters in the CSV file. In total 29.4%
334 of the clusters (213,367) were retained in the filtered CSV file (Table S7). Table 2 shows the
335 target- and non-target read counts of the monoculture samples after monoculture-based cluster
336 filtering. Target read counts ranged from 92.78% to 99.95%, the per species averaged non-
337 target read counts ranged from 0.00% to 0.60%. The combined effect of the BLASTN and
338 monoculture-based cluster filtering on the mock-mixture root samples is evaluated at the end
339 of the result section of process 5b.

340   Process 5a. msGBS non-calibrated analysis. Figure 4AC show that high correlations,
341 ranging from R^2=0.72 to 0.94, between the within-species abundance and the biomass-based
342 species proportions were found. However, the wide range of slopes (ranging from 0.20 to 1.64)
343 show that the assessment of across-species abundance within mixed root samples is impossible
344 without proper calibration. The msGBS FPS in non-calibrated mode was 0.46% (Table S8). For
345 pool 1 and 2 no msGBS FNS were found in non-calibrated mode.

346   Process 5b. msGBS calibrated analysis. Figure 4BD illustrates the effect of the
347 calibration procedure; the correction for per species typical read yield correct the slope towards
348 1 while high correlations, ranging from R^2=0.85 to 0.98, were retained. The calibration key,
349 needed for this correction, was calculated from a set of 10 replicate calibration samples for each
350 species pool. Indeed, there was large variation in between species read counts per unit of root
351 biomass (Table S9A and S10A). The averaged read counts varied from 609 to 13,298 in pool 1
352 and 315 to 4,597 in pool 2. This again illustrates why across-abundances within root samples
353 cannot be based directly on read counts. As in the qPCR method of Mommer et al 2008, the

calibration procedure is sensitive for signal variation between the calibration samples, due to errors in weighing equal tiny fresh biomasses by hand. Specifically, msGBS also requires comparable relative read counts between the calibration samples (Table S9B and S10B). Outlier values, containing one or more values that deviated more than 2.5STD, were detected (sample 1 of pool 1 and sample 15 and 18 of pool 2) and removed (Table S9B and S10B, Fig. S5 and S6). Removal of outlier 'calibration' samples is standard procedure in the old qPCR method and justified because of the sensitive root weighing procedure.

The biomass-based species proportions of the pool 1 and 2 mock-mixture samples was compared to the msGBS- and qPCR across-species abundances (*sensu* Mommer et al. 2008)(Fig. 5AB). In general, we found comparable correlations ranging from $R^2=0.84$ to 0.97 for msGBS and $R^2=0.94$ to 0.98 for qPCR (Table S11 and S12). These high correlations show that msGBS in calibrated mode can reliably estimate the across-species abundance within root mixtures. The slopes of the regression models ranged from 0.64 to 1.04 for msGBS and from 0.81 to 1.14 for qPCR. For 8 out of 16 species the msGBS slopes were within the 0.95 confidence interval boundaries of the slope=1. The regression models can be used to estimate across-species abundance from the msGBS signal of unknown experimental samples. Further analysis indicated that msGBS average FPS (0.88%) of the mock samples are comparable to those obtained using qPCR (0.43%)(Table S13 and S14). In pool 2 *Plantago lanceolata* and *Holcus lanatus* had a relative high average msGBS FPS (4.6% and 3.5%, respectively). For pool 1 and 2 no msGBS FNS were found in calibrated mode. The msGBS FPS in calibrated mode were comparable to those in non-calibrated mode (0.42% and 0.46%, respectively, Table S8).

The effect of cluster filtering

The combined effect of the BLASTN- and monoculture-based cluster filtering on the msGBS results was evaluated by comparing the msGBS and qPCR across-species abundances of the mock-mixture samples of pool 1, with- and without the combined filtering steps (Table S15). The average correlation (R^2) of the across-species abundance and biomass-based species proportions was improved from 0.95 to 0.98 by the combined optimizations. The average FPS was lowered from 2.05% to 0.42% percent by the combined optimizations. This demonstrates the effectivity of the BLASTN- and monoculture-based filtering in quenching false positive signals.

386

The influence of species assembly on the calibration key

To assess the influence of species assembly on the calibration key values and the regression model slopes of the biomass-based species proportions and non-calibrated msGBS estimated within-species abundance we compared these values of the three species present in both pools (Fig. S7). Due to the species assembly the slope shift ranged from 0.01 to 0.03, the calibration key shift ranged from 0.01 to 0.04 (Table S16).

Dutch field study msGBS results

On average 44,044 meta-reference clusters were generated per monoculture leaf sample; on average 1 cluster per 76 PE reads, more reads compared to the 1 cluster per 12 PE reads of *the Jena study* monocultures. This can be explained by two reasons 1) the Vsearch minuniquesize parameter which was increased from 2 to 3 because of the higher per monoculture sequence read input and 2) the fact that the Jena study monocultures derived from root, rather than leaf material which is expected to contain more Bacteria and Fungi. The second reason is hypothesized to be the cause of the difference in the percentage of leaf monoculture- and root field sample assembled reads retained in the filtered CSV file; 18% and 6%, respectively (Table S4).

The non-calibrated msGBS data of the *Dutch field study* was used to evaluate the congener specificity; does the monoculture based cluster filtering effectively identify between congener species homologous clusters and sufficiently increase the taxonomic resolution? To answer this question we looked at the msGBS rFPS of 'absent' species (based on extensive field surveys), which is assumed to be caused by the presence of congener species. This rFPS is calculated by dividing the msGBS signal of the 'absent' congener specie(s) by the msGBS signal of the 'present' congener specie(s). The analysis was based on 5 congener pairs, 5 congener triplets and 1 congener quartet. For congener triplets and quartets comparisons were performed in all available combinations (Table 3). For 5 congener combinations we found less than 0.5% rFPS and for 3 congener combinations between 0.5% and 3% rFPS. For the remaining 3 congener combinations we found a higher rFPS ranging from 14.53% to 43.96%. A closer look at the actual field survey data and their msGBS within-species abundances was used to discus these high rFPS signals (Fig. S8 and Fig. S9).

Discussion

As one of the very few molecular techniques to quantify relative species abundance in mixed root samples, the qPCR method of Mommer et al. (2008) produces robust results but also has its limitations. Here we present a new molecular method that solves these drawbacks by 1) allowing analysis of, essentially, an unlimited number of species in a single root sample, 2) the increased sensitivity to low species abundances compared to qPCR due to scalable sequencing effort, 3) the labor friendliness, 4) the prevention of PCR bias due to the use of Unique Molecule Identifiers (UMI's) and 5) the relative low laboratory costs (32 euro per sample). Our analysis show that msGBS is a very robust high-throughput molecular method to quantify across-species abundance related to root biomass (in calibrated mode) or within-species abundance across samples (in non-calibrated mode) in mixed root samples. Results of msGBS and qPCR were highly correlated in calibrated mode. msGBS had no False Negative Signals (FNS) and low (relative) False Positive Signals ((r)FPS) in most cases showing unprecedented taxonomic resolution. Out of 11 congener comparisons, only between two very closely related congener pairs significant rFPS was reported. The msGBS labwork is slightly more technical but more affordable compared to the current state of the art (RevMet).

msGBS thus outperforms other techniques on taxonomic resolution although extensive tests are yet to be performed and improvements with other techniques are possible. The taxonomic resolution, at congener level, is insufficiently validated for the currently available DNA based techniques (qPCR (Mommer et al., 2008), metabarcoding (Matesanz et al., 2019) and shotgun metabarcoding (Lang et al., 2019)). For RevMet (Peel et. al, 2019) only limited congener data is available but so far high rFPS are reported for one of two tested congener pairs. Smart application of filtering strategies or MinHash (Ondov et al., 2019) based analysis might, in the near future, further increase the taxonomic resolution of HTS based sequence data independent of origin of the used sequences (e.g. GBS, genome skimming or MinION), especially when more NCBI RefSeq genomes become available.

Methodological considerations regarding msGBS

msGBS library preparation

msGBS libraries were prepared for all 120 species of both *Jena* and *Dutch field Study*. The observed variation in number of demultiplexed sequence reads between samples is not

450   uncommon for GBS based techniques (Gardner et al., 2014; Sonah et al., 2013), and is

451   suggested to be the result of variation in gDNA quality (especially the presence of secondary

452   metabolites and ethanol residues). qPCR based msGBS library pooling, as performed for *Dutch*

453   *field Study* samples, accommodated more balanced sequencing output. Over all our advice is to

454   aim for 3M PE sequence reads for all sample types which results in sufficient meta-reference

455   clusters, an efficient monoculture-based cluster filtering, proper calibration and robust

456   estimation of species abundances of the mock-mixture- and unknown experimental samples.

457   Using qPCR based pooling allows for 120-140 samples to be processed in a single Hiseq X-

458   Ten sequence lane.

459

460   *Meta-reference assembly*

461   The number of GBS reference clusters generated depends on the sequencing effort, the

462   restriction enzyme choice, clustering parameters and genome related properties of a species. In

463   msGBS, the restriction enzyme choice cannot be optimized per species. For the *Jena field study*

464   data we observed a high variation in the number of clusters generated per species ranging from

465   8,475 for *Geranium pratense* to 260,885 for *Phleum pratense*. However, this was strongly

466   correlated to the sequencing effort. For the *Dutch field study,* were we aimed for 3M sequence

467   reads per sample, we observed much less between species variation in cluster numbers. Despite

468   the large variation, sufficient clusters were yielded for the *Jena field study* to allow robust

469   estimation of the across- or within-species abundance. Overall, we do not regard cluster number

470   variation as a fundamental problem since the mock-mixture reads are all mapped to the same

471   set of clusters.

472

473   *BLASTN filtering*

474   The processing of the *Jena field study* meta-reference BLASTN output led to the

475   removal of only 1% of the clusters. Removed clusters were predominantly annotated to AM

476   Fungi and Bacteria (60% and 34%, respectively). Many clusters could not be identified because

477   of the incompleteness of the NCBI nr database used. A demonstration of this is that >99% of

478   the removed AM Fungi clusters had a hit against *Rhizophagus irregularis* strain

479   DAOM_181602=DAOM_197198; the only AM Fungi (Tisserant et al., 2013) of which

480   genome-scale sequence information is present in the NCBI nr database. The monoculture

481 material of the *Dutch field study*, for meta-reference assembly, was collected from aboveground
482 leaf material to prevent unnecessary interference with soil biota.
483
484 *Mapping*
485   We used assembled (merged and joined) reads for mapping instead non-assembled
486 reads. We believe that, for msGBS, assembled reads is preferable; some fragments are in the
487 size region were 20 bp overlap, needed for a read to merge, is just present for some read pairs
488 but not for others resulting in merged and joined variants of the same locus. During clustering
489 those variants are not collapsed and therefor result in more than one cluster. The mapping of
490 assembled reads prevents bias as they will only map to either the merged or non-merged variant
491 cluster of that locus.
492
493 *Monoculture-based cluster filtering*
494   Evaluation of the per cluster read counts in the CSV file showed that it was quite
495 common that monoculture reads of multiple species were mapped to a single meta-reference
496 cluster. Monoculture-based cluster filtering identifies clusters with relative high non-target
497 mapping. Non-target mapped reads can be caused by 1) between species homologous clusters
498 2) clusters that originated from root- or rhizosphere microbiota 3) non-target roots present in
499 the monoculture plots or laboratory environment pollution and 4) tag- or index jumping
500 (Schnell et al., 2015) although this is mainly a problem in library types that have blunt-end
501 ligation steps in the wet protocol. Oram et al. (2018) reported that the *Jena field study*
502 monoculture material of *Holcus lanatus* and *Poa pratensis* contained traces of *Plantago*
503 *lanceolata*. Due to the monoculture-based cluster filtering, we found no significant elevated
504 signal for these species using msGBS. The monoculture-based cluster filtering lowered the FPS.
505 However, a low number of clusters (e.g. in *Geranium pratense)* will cause the monoculture-
506 based cluster filtering to be less effective; the detection of between species homologous clusters
507 is only possible when those clusters are present. As a consequence mock-mixture sample *G.*
508 *pratense* reads which are not represented in the *G. pratense* meta-reference cluster set, might
509 map to other species clusters causing a higher FPS in those species. This might explain the
510 higher average FPS reported for pool 2.
511   For the *Jena field study*, we accepted a maximum of 6.7% (f2=15) non-target reads
512 resulting in high between msGBS and qPCR correlations, acceptable FPS, no FNS and minimal

513   sample loss due to filter f3 (1000 reads). For the *Dutch field study*, were we had a higher

514   monoculture sample read average, we accepted a maximum of 0.33% (f2=300) non-target reads

515   resulting in low rFPS and only two discarded samples due to filter f3 (2000 reads).

516

517   *msGBS in non-calibrated and calibrated mode*

518   Our results showed that msGBS in non-calibrated mode resulted in slightly lower

519   correlations between biomass-based species proportions and within-species abundances

520   compared msGBS in calibrated mode. msGBS in both non-calibrated and mode resulted in low

521   mock-mixture sample FPS . In general the msGBS results of pool 1 were more robust compared

522   to those of pool 2. We believe this was mainly due to a lower on average sequence effort for

523   the pool 2 samples; especially the insufficient sequencing effort of the pool 2 *Geranium*

524   *pratense* monoculture sample and thus the low number of *Geranium pratense* clusters in the

525   meta-reference which is hypothesized to result in a less efficient monoculture-based cluster

526   filtering and higher FPS in the non-target species.

527   msGBS in calibrated mode delivered results comparable to the qPCR-based method of

528   Mommer et al. (2008). The Calibration procedure was able to correct for the 22-fold differences

529   in, across-species, read mapping counts in the calibration samples of pool 1. Some variation in

530   per species relative sequence read mapping counts was observed between calibration samples.

531   This variation is likely due to the small amounts of root biomass used per species in these

532   samples (12.5 mg per species; 8 species), the manually weighing procedure where differences

533   in root morphology and moistness of the monoculture roots created errors. The use of replicate

534   calibration samples enables the removal of outliers to ensure the calculation of a representable

535   'calibration key'. Overall, the across-species mapping counts between calibration samples were

536   stable within species pools which was a prerequisite for the msGBS in calibrated mode.

537

538   *False positive and false negative signals (FPS, FNS)*

539   No FNS was detected within the *Jena field study* mock-mixture samples in both msGBS

540   mode. The average FPS of the qPCR and msGBS calibrated data were similar for pool 1. For

541   pool 2 a relative high FPS was recorded for *Plantago lanceolata* and *Holcus lanatus* possibly

542   partly due to the low sequencing effort of the *Geranium pratense* monoculture root sample as

543   discussed above. But this is at least partly contradicted by the fact that there was, for both

544   species, a high variation in FPS between samples which directs more to pollution of

545 monoculture field plots. For the analysis of experimental samples of unknown composition low

546 FPS rates are important. The low FPS rates (<1%) observed for pool 1 are acceptable for

547 analysis of field samples; an analytical detection limit of 1% can be introduced. But, based on

548 the results of the *Dutch field study,* we believe msGBS can perform even better with a higher

549 sequencing effort and the use of leaf material for meta-reference assembly. The FPS analysis

550 of the *Dutch field study* cannot be executed in the way of the *Jena field study*; the biomass-

551 based abundances of the root samples are unknown. We used the msGBS signals of congener

552 species to review the taxonomic resolution in terms of relative FPS (rFPS) as discussed below.

553

554 *msGBS taxonomic resolution*

555   Taxonomic resolution is an unresolved issue in plant taxonomy studies due to high

556 homologies between closely related congener species and is further complicated by a plethora

557 of natural hybrids. The use of longer sequences can solve this issue but current long read

558 sequencers do not deliver premium quality reads nor sufficiently read numbers. When using

559 huge numbers of smaller but high quality Hiseq reads many assembled meta-reference clusters

560 are highly homologous between species. The abundant presence of Bacterial and Fungi in plant

561 roots further complicate species-specific quantification.

562   The effect of monoculture-based cluster filtering, which identifies between species

563 homologous clusters, on the taxonomic resolution is best evaluated at congener species level.

564 Of all current available techniques that target plant material, msGBS is best compared to

565 RevMet (Peel et al., 2019). To our knowledge, RevMet and msGBS are the only HTS based

566 method that uses non-targeted sequencing data (and not an extracted metabarcoding or

567 mitogenome subset) for the quantification of plant species relative abundances within mixed

568 species samples.

569   The RevMet mapping data of the 2 congener pairs (*Papaver* and *Ranunculus*; REF)

570 present in their mock-mixture dataset was used to calculate the rFPS. The rFPS of RevMet and

571 the *Dutch field study* msGBS data were compared to evaluate the taxonomic resolution of both

572 methods (Table 2). For one of the two congener pairs of RevMet, and for 3 of the 11 congener

573 groups of msGBS a high rFPS (>3%) was recorded. For 2 of the 3 high rFPS msGBS cases, the

574 interference was consistent over samples and within a single pair of species, corresponding with

575 close phylogenetic relatedness. Within the *Ranunculus* congener triplet, visual inspection (Fig.

576 S8) of the *Ranunculus* msGBS signals confirmed that the rFPS within *R. repens* was solely

577 caused by *R. bulbosus* sequence reads. No interference between the *R. acris* and both *R. repens*
578 or *R. bulbosus* signals was observed, this corresponded to their phylogenetic relatedness
579 (Baltisberger & Hörandl, 2016). Within the *Rumex* congener triplet; visual inspection (Fig. S9)
580 of the *R. thyrsiflorus* and *R. acetose* signal showed interference in both directions. For *R. crispus*
581 no significant rFPS was detected from the other two species, this again corresponds to their
582 phylogenetic relatedness (Schuster, Reveal, Bayly, & Kron, 2015). More RevMet congener data
583 is needed to properly evaluate the RevMet taxonomic resolution but in the single case were a
584 direct comparison with msGBS could be made (rFPS between *R. repens* and *R. acris*) the latter
585 showed an improved taxonomic resolution.
586

587 *Genetic variation and natural hybrids*
588 Genetic variation within species may be another source of error. Theoretically, genetic
589 variation can cause bias through 1) variable PCR efficiencies caused by mutations in primer-
590 or restriction enzyme binding sites or 2) erroneous identification. This might also be the case
591 in msGBS. For msGBS analysis at species level we expect very limited genetic bias because of
592 four reasons; 1) the use of thousands of clusters per species 2) the read mapping is based on
593 95% identity 3) the application of monoculture-based cluster filtering and 4) the fact that we
594 use universal primers which, during PCR, anneal to the ligated adapters and not to the genetic
595 variable gDNA sequence itself.
596 Within congener groups natural hybrids are hypothesized to cause a msGBS signal for
597 both hybrid donor species. This signal can be misinterpreted as FPS when the hybrid is falsely
598 classified as either donor species during field survey. Especially during the collection of
599 monoculture plant material, one must be cautious of hybrids.
600

601 **msGBS application on multi-species samples**
602

603 The origin of the monoculture roots, from which the calibration samples were
604 assembled, is important for correct calibration of mock-mixture- and unknown experimental
605 samples. The environment in which the monoculture roots are harvested should be similar to
606 the experimental conditions with regard to soil type, growth conditions and plant age. Species
607 pool slightly affected the relative sequence read mapping counts of individual species. This was
608 demonstrated by comparing the calibration key values of the three species present in both pools

609 to the slope of the regression model of the biomass-based species proportions and non-
610 calibrated msGBS estimated per species abundance. For optimal calibration it is advisable to
611 produce calibration samples separate for each experimental condition and timepoint (season,
612 year). To minimize the chance that, in a natural field setting, species are missing in the meta-
613 reference the monoculture material for the creation of the meta-reference is best collected
614 aboveground and throughout the year. Species with latent presence in the field plot in the form
615 of seeds or tubers will not interfere with the msGBS signals as roots are first washed from the
616 soil core. Significant FPS signals from missing species are only expected when the species are
617 present in the form of roots and when closely related to species in the meta-reference.

618

619   Sampling representative pure species-specific fine root tissue in high diversity plant
620 communities in natural field settings will often be difficult to impossible. We have shown that
621 even without the preferred calibration, msGBS can provide meaningful results on quantitative
622 distribution differences for the species in the plant community. For example, the within-species
623 relative to total sequence read mapping counts can be compared between samples of different
624 locations and soil depth. In this way, the distribution of roots of a single species in the soil
625 column can be compared to soil type, soil heterogeneity and the presence of other species.
626 Likewise, the degree of clustering of roots in the horizontal plane may unravel spatial niches
627 belowground that cannot be derived from aboveground patterns because roots generally have a
628 much wider range than shoots. Although root quantities cannot be compared between species,
629 root distributions can, by which positive or negative associations may be unraveled related to
630 questions of species competition and facilitation. These new opportunities for studying
631 belowground community assembly in relation to environmental change now open up even for
632 most diverse plant communities such a species-rich grasslands (Frank et al., 2010; Kesanakurti
633 et al., 2011) and tropical forests (Jones et al., 2011).

634

635 Conclusion
636
637   Our results highlight msGBS in calibrated mode as a novel, robust and cost-effective
638 approach to estimate across-species abundances in mixed root samples. We showed that
639 msGBS can as well be used in non-calibrated mode to estimate within-species abundances in
640 high diversity plant communities when the arduous assembly of calibration samples is not
641 preferred. msGBS has a high taxonomic resolution and is well able to distinguish congener

species. However, the genetic distance between closely related congener species approaches to the within-species genetic distance and the genetic gap is in some cases filled by a spectrum of hybrid variants. Although msGBS was developed with plant roots in mind the methodology is applicable to other sample types like pollen- or diatom mixtures.

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, *9*(1), 134–147. https://doi.org/10.1111/2041-210X.12849

Baltisberger, M., & Hörandl, E. (2016). Karyotype evolution supports the molecular phylogeny in the genus Ranunculus (Ranunculaceae). *Perspectives in Plant Ecology, Evolution and Systematics*, *18*, 1–14. https://doi.org/10.1016/j.ppees.2015.11.001

Barry, K. E., Weigelt, A., van Ruijven, J., de Kroon, H., Ebeling, A., Eisenhauer, N., … Mommer, L. (2019). Above- and belowground overyielding are related at the community and species level in a grassland biodiversity experiment. *Advances in Ecological Research*. https://doi.org/10.1016/BS.AECR.2019.05.001

Bobowski, B., Hole, D., Wolf, P., & Bryant, L. (1999). Identification of roots of woody species using polymerase chain reaction (PCR) and restriction fragment length polymorphism (RFLP) analysis. *Molecular Ecology*, *8*(3), 485–491. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10199009

Bramham, L., & Pyke, K. (2017). Changing plastid dynamics within early root and shoot apical meristem-derived tissue of A. thaliana. *Bioscience Horizons*, *10*, 1–10. https://doi.org/10.1093/biohorizons/hzx001

Braun-blanquet, J. (1932). *Plant sociology. The study of plant communities. First ed.* McGraw-Hill Book Co., Inc., New York and London.

Brundrett, M. (2004). Diversity and classification of mycorrhizal associations. *Biological Reviews*, *79*(3), 473–495. https://doi.org/DOI: 10.1017/S1464793103006316

679 Brunner, I., Brodbeck, S., Büchler, U., & Sperisen, C. (2001). Molecular identification of fine
680   roots of trees from the Alps: Reliable and fast DNA extraction and PCR-RFLP analyses of
681   plastid DNA. *Molecular Ecology*, *10*(8), 2079–2087. https://doi.org/10.1046/j.1365-
682   294X.2001.01325.x

683 Ebeling, A., Pompe, S., Baade, J., Eisenhauer, N., Hillebrand, H., Proulx, R., … Weisser, W. W.
684   (2014). A trait-based experimental approach to understand the mechanisms underlying
685   biodiversity-ecosystem functioning relationships. *Basic and Applied Ecology*, *15*(3), 229–
686   240. https://doi.org/10.1016/j.baae.2014.02.003

687 Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E.
688   (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity
689   species. *PLoS ONE*, *6*(5), 1–10. https://doi.org/10.1371/journal.pone.0019379

690 Frank, D. A., Pontes, A. W., Maine, E. M., Caruana, J., Raina, R., Raina, S., & Fridley, J. D.
691   (2010). Grassland root communities: Species distributions and how they are linked to
692   aboveground abundance. *Ecology*, *91*(11), 3201–3209. https://doi.org/10.1890/09-
693   1831.1

694 Frank, D. A., Pontes, A. W., Maine, E. M., & Fridley, J. D. (2015). Fine-scale belowground
695   species associations in temperate grassland. *Molecular Ecology*, *24*(12), 3206–3216.
696   https://doi.org/10.1111/mec.13232

697 Gardner, K. M., Brown, P., Cooke, T. F., Cann, S., Costa, F., Bustamante, C., … Myles, S. (2014).
698   Fast and Cost-Effective Genetic Mapping in Apple Using Next-Generation Sequencing.
699   *G3&#58; Genes|Genomes|Genetics*, *4*(9), 1681–1687.
700   https://doi.org/10.1534/g3.114.011023

701 Garrido-Sanz, L., Senar, M. À., & Piñol, J. (2020). Estimation of the relative abundance of
702   species in artificial mixtures of insects using low-coverage shotgun metagenomics.
703   *Metabarcoding and Metagenomics*, *4*, 1–18. https://doi.org/10.3897/mbmg.4.48281

704 Hendriks, M., Ravenek, J. M., Smit-Tiekstra, A. E., van der Paauw, J. W., de Caluwe, H., van der
705   Putten, W. H., … Mommer, L. (2015). Spatial heterogeneity of plant-soil feedback affects
706   root interactions and interspecific competition. *New Phytologist*, *207*(3), 830–840.
707   https://doi.org/10.1111/nph.13394

708 Hiiesalu, I., Öpik, M., Metsis, M., Lilje, L., Davison, J., Vasar, M., … Pärtel, M. (2012). Plant
709   species richness belowground: Higher richness and new patterns revealed by next-
710   generation sequencing. *Molecular Ecology*, *21*(8), 2004–2016.
711   https://doi.org/10.1111/j.1365-294X.2011.05390.x

712 Jackson, R. B., Moore, L. A., Hoffmann, W. A., Pockman, W. T., & Linder, C. R. (1999).
713   Ecosystem rooting depth determined with caves and DNA. *Proceedings of the National
714   Academy of Sciences*, *96*(20), 11387–11392. https://doi.org/10.1073/pnas.96.20.11387

715 Jones, F. A., Erickson, D. L., Bernal, M. A., Bermingham, E., Kress, W. J., Herre, E. A., … Turner,

716   B. L. (2011). The roots of diversity: Below ground species richness and rooting
717   distributions in a tropical forest revealed by DNA barcodes and inverse modeling. *PLoS*
718   *ONE*, *6*(9). https://doi.org/10.1371/journal.pone.0024506

719   Kesanakurti, P. R., Fazekas, A. J., Burgess, K. S., Percy, D. M., Newmaster, S. G., Graham, S. W.,
720   … Husband, B. C. (2011). Spatial patterns of plant diversity below-ground as revealed by
721   DNA barcoding. *Molecular Ecology*, *20*(6), 1289–1302. https://doi.org/10.1111/j.1365-
722   294X.2010.04989.x

723   Lang, D., Tang, M., Hu, J., & Zhou, X. (2019). Genome-skimming provides accurate
724   quantification for pollen mixtures. *Molecular Ecology Resources*, *19*(6), 1433–1446.
725   https://doi.org/10.1111/1755-0998.13061

726   Leidenfrost, R. M., Bänsch, S., Prudnikow, L., Brenig, B., Westphal, C., & Wünschiers, R. (2020).
727   Analyzing the Dietary Diary of Bumble Bee. *Frontiers in Plant Science*, *11*(March), 1–9.
728   https://doi.org/10.3389/fpls.2020.00287

729   Linder, C. R., Moore, L. A., & Jackson, R. B. (2000). A universal molecular method for
730   identifying underground plant parts to species. *Molecular Ecology*, *9*(10), 1549–1559.
731   https://doi.org/10.1046/j.1365-294X.2000.01034.x

732   Matesanz, S., Pescador, D. S., Pías, B., Sánchez, A. M., Chacón-Labella, J., Illuminati, A., …
733   Escudero, A. (2019). Estimating belowground plant abundance with DNA metabarcoding.
734   *Molecular Ecology Resources*, *19*(5), 1265–1277. https://doi.org/10.1111/1755-
735   0998.13049

736   Mckay, A., Riley, I. T., Hartley, D., Wiebkin, S., Li, G., Coventry, S., & Huang, C. (2008). Studying
737   root development in soil using DNA technology : idea to impact. *Society*, (September).

738   McNickle, G. G., Cahill, J. F., & Deyholos, M. K. (2008). A PCR-based method for the
739   identification of the roots of 10 co-occurring grassland species in mesocosm
740   experiments. *Botany*, *86*(5), 485–490. https://doi.org/10.1139/B08-014

741   Mommer, L., Wagemaker, C. A. M., De Kroon, H., & Ouborg, N. J. (2008). Unravelling below-
742   ground plant distributions: A real-time polymerase chain reaction method for
743   quantifying species proportions in mixed root samples. *Molecular Ecology Resources*,
744   *8*(5), 947–953. https://doi.org/10.1111/j.1755-0998.2008.02130.x

745   Mommer, Liesje, Dumbrell, A. J., Wagemaker, C. A. M., & Ouborg, N. J. (2011). Belowground
746   DNA-based techniques: Untangling the network of plant root interactions. *Plant and Soil*,
747   *348*(1–2), 115–121. https://doi.org/10.1007/s11104-011-0962-0

748   Mommer, Liesje, van Ruijven, J., de Caluwe, H., Smit-Tiekstra, A. E., Wagemaker, C. A. M.,
749   Joop Ouborg, N., … de Kroon, H. (2010). Unveiling below-ground species abundance in a
750   biodiversity experiment: A test of vertical niche differentiation among grassland species.
751   *Journal of Ecology*, *98*(5), 1117–1127. https://doi.org/10.1111/j.1365-
752   2745.2010.01702.x

753 Ondov, B. D., Starrett, G. J., Sappington, A., Kostic, A., Koren, S., Buck, C. B., & Phillippy, A. M.
754   (2019). Mash Screen: High-throughput sequence containment estimation for genome
755   discovery. *Genome Biology*, *20*(1), 1–13. https://doi.org/10.1186/s13059-019-1841-x

756 Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., &
757   Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using
758   MinHash. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/s13059-016-0997-x

759 Oram, N. J., Ravenek, J. M., Barry, K. E., Weigelt, A., Chen, H., Gessler, A., … Mommer, L.
760   (2018). Below-ground complementarity effects in a grassland biodiversity experiment
761   are related to deep-rooting species. *Journal of Ecology*, *106*(1), 265–277.
762   https://doi.org/10.1111/1365-2745.12877

763 Padilla, F. M., Mommer, L., de Caluwe, H., Smit-Tiekstra, A. E., Wagemaker, C. A. M., Ouborg,
764   N. J., & de Kroon, H. (2013). Early Root Overproduction Not Triggered by Nutrients
765   Decisive for Competitive Success Belowground. *PLoS ONE*, *8*(1), 1–9.
766   https://doi.org/10.1371/journal.pone.0055805

767 Peel, N., Dicks, L. V., Clark, M. D., Heavens, D., Percival-Alwyn, L., Cooper, C., … Yu, D. W.
768   (2019). Semi-quantitative characterisation of mixed pollen samples using MinION
769   sequencing and Reverse Metagenomics (RevMet). *Methods in Ecology and Evolution*,
770   *10*(10), 1690–1701. https://doi.org/10.1111/2041-210X.13265

771 Ridgway, K. P., Duck, J. M., & Peter W Young, J. (2003). Identification of roots from grass
772   swards using PCR-RFLP and FFLP of the plastid trnL (UAA) intron. *BMC Ecology*, *3*, 6–11.
773   https://doi.org/10.1186/1472-6785-3-8

774 Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated - reducing
775   sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology*
776   *Resources*, *15*(6), 1289–1303. https://doi.org/10.1111/1755-0998.12402

777 Schuster, T. M., Reveal, J. L., Bayly, M. J., & Kron, K. A. (2015). An updated molecular
778   phylogeny of Polygonoideae (Polygonaceae): Relationships of Oxygonum, Pteroxygonum,
779   and Rumex, and a new circumscription of Koenigia. *Taxon*, *64*(6), 1188–1208.
780   https://doi.org/10.12705/646.5

781 Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., … Belzile, F. (2013). An
782   Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and
783   Efficiency of SNP Discovery and Genotyping. *PLoS ONE*, *8*(1), 1–9.
784   https://doi.org/10.1371/journal.pone.0054603

785 Suhl, J. A., Chopra, P., Anderson, B. R., Bassell, G. J., & Warren, S. T. (2014). R: A Language and
786   Environment for Statistical Computing. *Nature*, Vol. 485, pp. 237–241.
787   https://doi.org/10.1093/hmg/ddu272

788 Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., … Willerslev, E.
789   (2007). Power and limitations of the chloroplast trnL (UAA) intron for plant DNA

790   barcoding. *Nucleic Acids Research*, *35*(3). https://doi.org/10.1093/nar/gkl938

791   Team), (Python core. (2015). *Python: A dynamic, open source programming language.* Python
792   Software Foundation.

793   Team), (RStudio. (2016). *RStudio: Integrated Development Environment for R*. Retrieved from
794   http://www.rstudio.com/

795   Tisserant, E., Malbreil, M., Kuo, A., Kohler, A., Symeonidi, A., Balestrini, R., … Martin, F. (2013).
796   Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant
797   symbiosis. *Proceedings of the National Academy of Sciences*, *110*(50), 20117–20122.
798   https://doi.org/10.1073/pnas.1313452110

799   Wildová, R. (2004). Below-ground spatial pattern of rhizomes in a grassland community and
800   its relevance to above-ground spatial pattern. *Plant Ecology Formerly `Vegetatio'*, *174*(2),
801   321–338. https://doi.org/10.1023/b:vege.0000049111.47743.14

802   Zeng, W., Xiang, W., Zhou, B., Lei, P., & Zeng, Y. (2017). Measurement of belowground
803   diversity of fine roots in subtropical forests based on a quantitative real-time PCR (qPCR)
804   method. *Plant and Soil*, *420*(1–2), 539–552. https://doi.org/10.1007/s11104-017-3402-y

805

Data Accessibility Statement

Author Contributions

813   C.W. designed the research, C.W. and A.S performed the lab work, T.G. M.P. and C.W.

814   contributed to analytical tools, C.W. and H.K. analyzed the data, C.W., H.K., E.V., L.M. and

815   A.W. wrote the manuscript.

Supplementary files

817   MER_Supplemental_Information_Wagemaker_et_al_2020.pdf
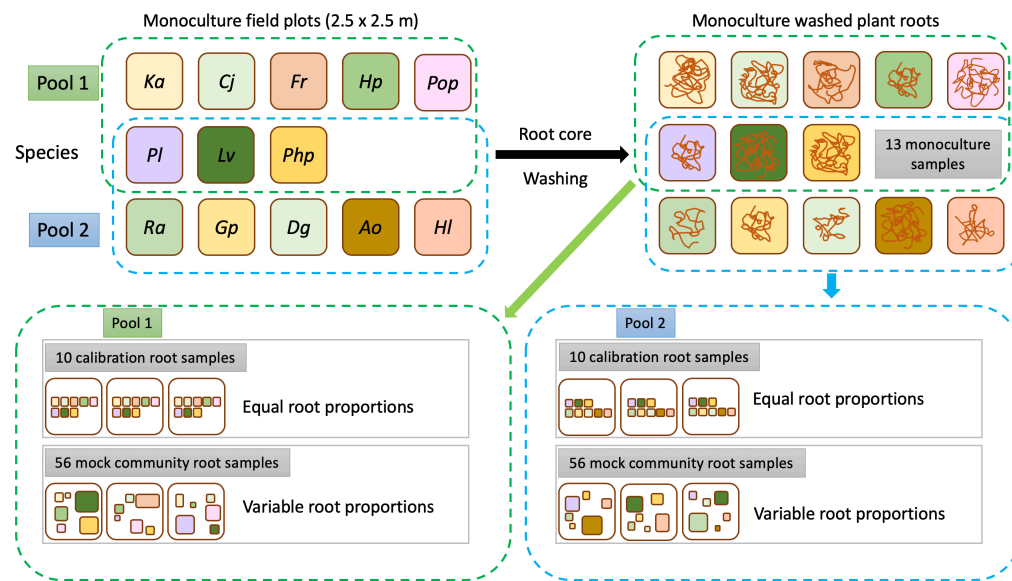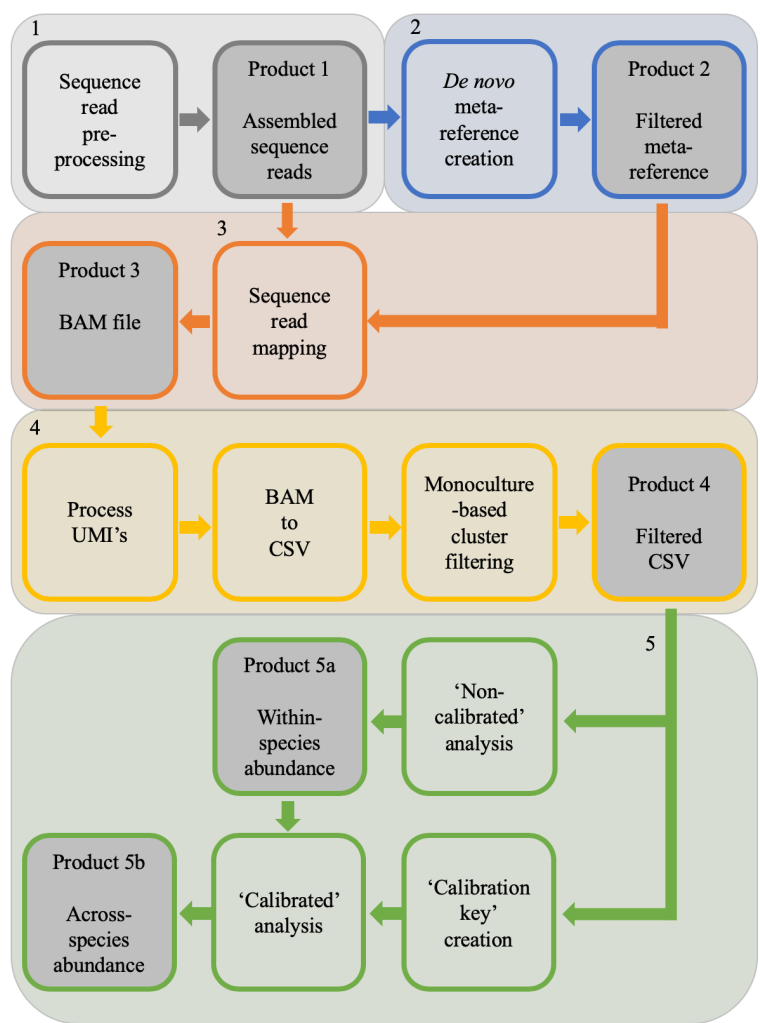
818

Tables and Figures

820

Figure 1 The *Jena field study* experimental setup. The monoculture, calibration and mock-mixture samples are assembled from washed monoculture root material from 2.5m x 2.5m monoculture field plots. Two species pools were created each consisting of 8 of the 13 species; these assemblies correlated to two 8-species field plots of the Jena experiment.

825

826

827

828　Figure 2 Overview of the msGBS analysis as outlined in the text and supporting documentation.
829　Process 1 (gray) depicts the pre-processing of the sequence reads and produces product 1; the
830　assembled sequence reads. Process 2 (blue) is the creation of the BLASTN filtered meta-
831　reference (product 2). Process 3 (orange) depict the sequence read mapping to produce a BAM
832　alignment file (product 3). Process 4 (yellow) is the identification of PCR duplicates, the
833　conversion of BAM to CSV format and the monoculture-based cluster filtering. Total, per
834　sample per cluster, read counts are stored in a filtered CSV file (product 4). Process 5 (green)
835　starts with the non-calibrated analysis which results in the within-species abundance (product
836　5a). Next, a calibration key was created from the calibration sample read counts. The calibration
837　key was subsequently used to convert the within-species abundance into the across-species
838　abundance (product 5b).

Monoculture read counts of *P. lanceolata* meta-reference cluster *Pl*_1

| Target monoculture read count |
| Non-target monoculture read counts |

Read counts	Cluster *Pl*_1	
	Reads	%
Total	1132	
Target	1094	100
Non-target	38	3.5

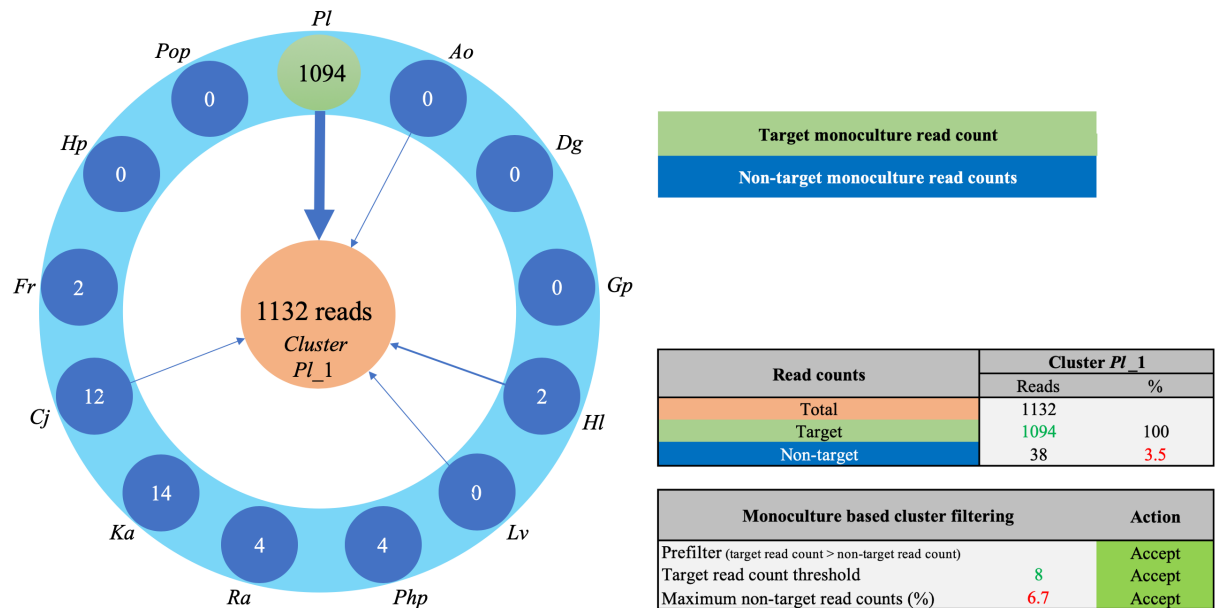| Monoculture based cluster filtering		Action
Prefilter (target read count > non-target read count)		Accept
Target read count threshold	8	Accept
Maximum non-target read counts (%)	6.7	Accept

Figure 3 Illustration of the monoculture-based cluster filtering (Fig. 2, process 4, step 3). The monoculture-based filtering evaluates the mapped read counts of the monoculture samples. Each cluster is evaluated individually. If a read from a monoculture sample was mapped to a cluster that originated from that monoculture sample this is called a target read count and if mapped to a cluster that originated from another monoculture sample this is called a non-target read count. In this example 1132 reads were counted for *Plantago lanceolata* meta-reference cluster *Pl*_1 which could be split in 1094 target read counts and 38 (3.5%) non-target read counts. The *Pl*_1 cluster did pass the all evaluation steps so this cluster was accepted and recorded in the filtered CSV file.

Pool 1	Species								
	P. lanceolata	*K. arvensis*	*C. jacea*	*F. rubra*	*H. pubescens*	*P. pratensis*	*L. vulgare*	*P. pratense*	
	Calibration sample								total
Biomass based proportions	1 :	1 :	1 :	1 :	1 :	1 :	1 :	1	
Averaged read counts	633 :	4429 :	2652 :	3119 :	3046 :	13137 :	2414 :	5147	34576
Calibration key	0.02 :	0.13 :	0.08 :	0.09 :	0.09 :	0.38 :	0.07 :	0.15	1
	Example mock mixture sample								total
Biomass based proportions	? :	? :	? :	? :	? :	? :	? :	?	
Read counts	247 :	2259 :	1121 :	4291 :	14 :	6520 :	546 :	894	15892
Between-species abundance	0.02 :	0.14 :	0.07 :	0.27 :	0.00 :	0.41 :	0.03 :	0.06	1
Calibrated msGBS signal	0.85 :	1.11 :	0.92 :	2.99 :	0.01 :	1.08 :	0.49 :	0.38	7.83
Across-species abundance	0.11 :	0.14 :	0.12 :	0.38 :	0.00 :	0.14 :	0.06 :	0.05	1

Table 1 Illustration of the final calculations of a mock-mixture sample in non-calibrated and calibrated mode. First the per species read counts of both calibration and mock-mixture samples are divided by total read count. The relative read counts of a mock-mixture sample is the 'within-species abundance' (green) in non-calibrated mode. For calibrated mode the averaged relative read counts of the calibration samples is called the calibration key (blue). In calibrated mode the across-species abundance of a mock-mixture sample is calculated in two steps. First the relative read counts of the mock-mixture sample is divided by the calibration key which results in the calibrated msGBS signal. The across-species abundance (red) is calculated by dividing the per species calibrated signal to the total calibrated signal.

Monoculture based cluster filtering parameters f1=8 f2=15 f3=2000		Species pool		Total meta-reference cluster read counts (%)													Average non-target read counts (%)	Total read counts
		1	2	*P. lanceolata*	*R. acris*	*K. arvensis*	*G. pratense*	*C. jacea*	*D. glomerata*	*A. odoratum*	*H. lanatus*	*F. rubra*	*H. pubescens*	*P. pratensis*	*L. vulgare*	*P. pratense*		
Monoculture sample	*P. lanceolata*	X	X	98.06	0.08	0.37	0.00	0.07	0.19	0.43	0.01	0.06	0.59	0.06	0.04	0.03	0.16	23536
	R. acris		X	0.18	99.65	0.05	0.00	0.01	0.00	0.01	0.02	0.00	0.00	0.05	0.01	0.02	0.03	64911
	K. arvensis	X		0.14	0.02	99.60	0.00	0.02	0.09	0.01	0.01	0.01	0.00	0.07	0.02	0.02	0.03	82920
	G. pratense		X	0.00	0.00	4.12	92.78	0.00	0.00	0.00	0.00	2.06	0.00	0.00	1.03	0.00	0.60	97
	C. jacea	X		0.19	0.04	0.13	0.00	98.52	0.02	0.00	0.09	0.19	0.05	0.49	0.09	0.19	0.12	22645
	D. glomerata		X	0.00	0.01	0.05	0.00	0.05	99.24	0.10	0.00	0.18	0.11	0.13	0.08	0.05	0.06	52738
	A. odoratum		X	0.01	0.03	0.01	0.00	0.04	0.01	99.58	0.06	0.13	0.02	0.01	0.08	0.02	0.04	11862
	H. lanatus		X	0.00	0.00	0.00	0.00	0.00	0.00	0.00	99.95	0.01	0.01	0.01	0.00	0.00	0.00	181146
	F. rubra	X		0.01	0.02	0.00	0.01	0.02	0.02	0.01	0.14	99.51	0.04	0.07	0.13	0.04	0.04	19868
	H. pubescens	X		0.01	0.01	0.01	0.00	0.00	1.14	0.01	0.07	0.04	98.57	0.03	0.03	0.08	0.12	14432
	P. pratensis	X		0.01	0.04	0.00	0.00	0.00	0.11	0.01	0.14	0.01	0.15	99.50	0.01	0.01	0.04	172862
	L. vulgare	X	X	0.00	0.03	0.06	0.00	0.03	0.00	0.06	0.17	0.04	0.07	0.06	99.48	0.00	0.04	6917
	P. pratense	X	X	0.02	0.05	0.02	0.00	0.01	0.04	0.04	0.03	0.06	0.03	0.09	0.11	99.49	0.04	116427

864 Table 2 Per species meta-reference cluster total read counts (%) of each monoculture sample
865 after monoculture-based cluster filtering. The read counts are expressed relative to the
866 monoculture sample total read count. The non-target signals are heatmap colored.
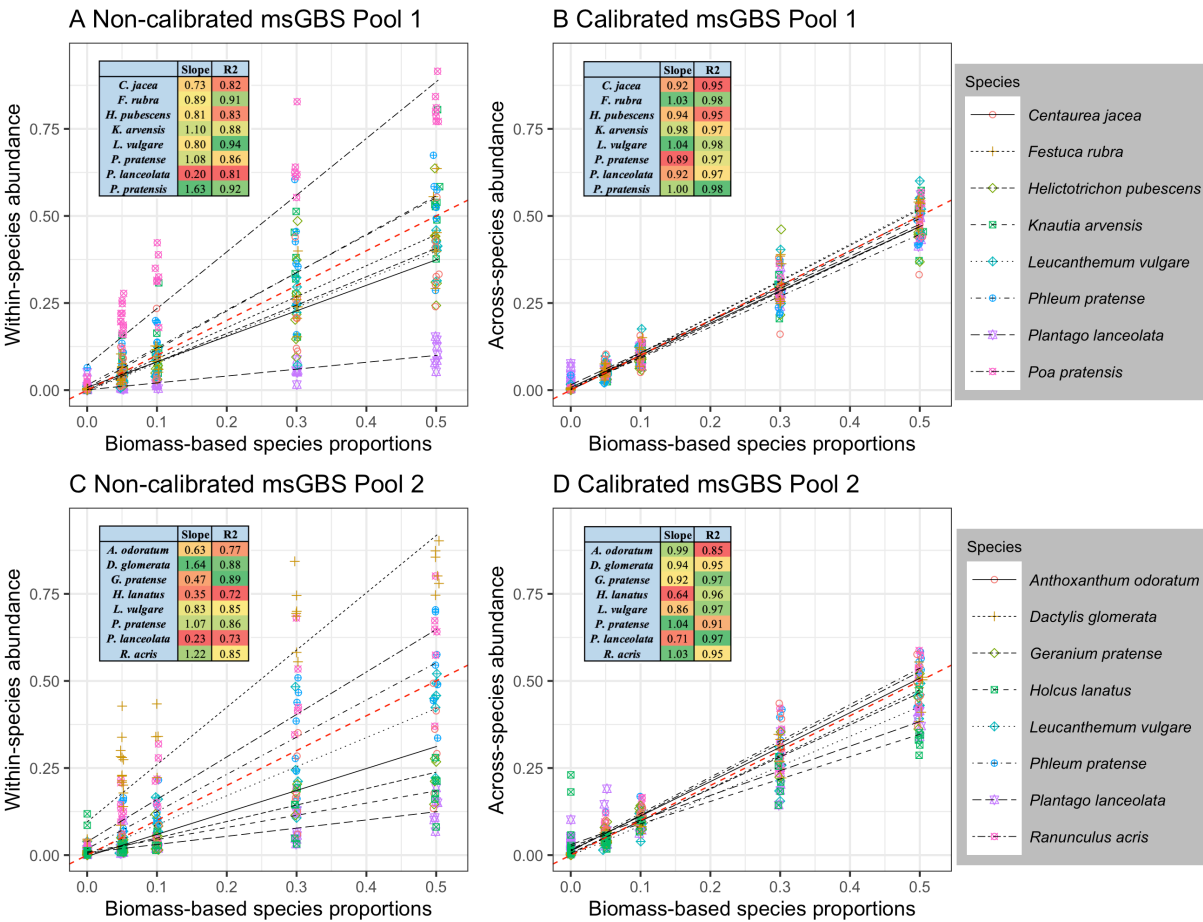867
868

869

870

871
872 Figure 4. Correlation of biomass-based species proportions to the msGBS- non-calibrated
873 within-species abundance and calibrated across-species abundance of the mock-mixture
874 samples of species pool 1 (AB) and 2 (CD). Regression line slopes and correlations are inserted
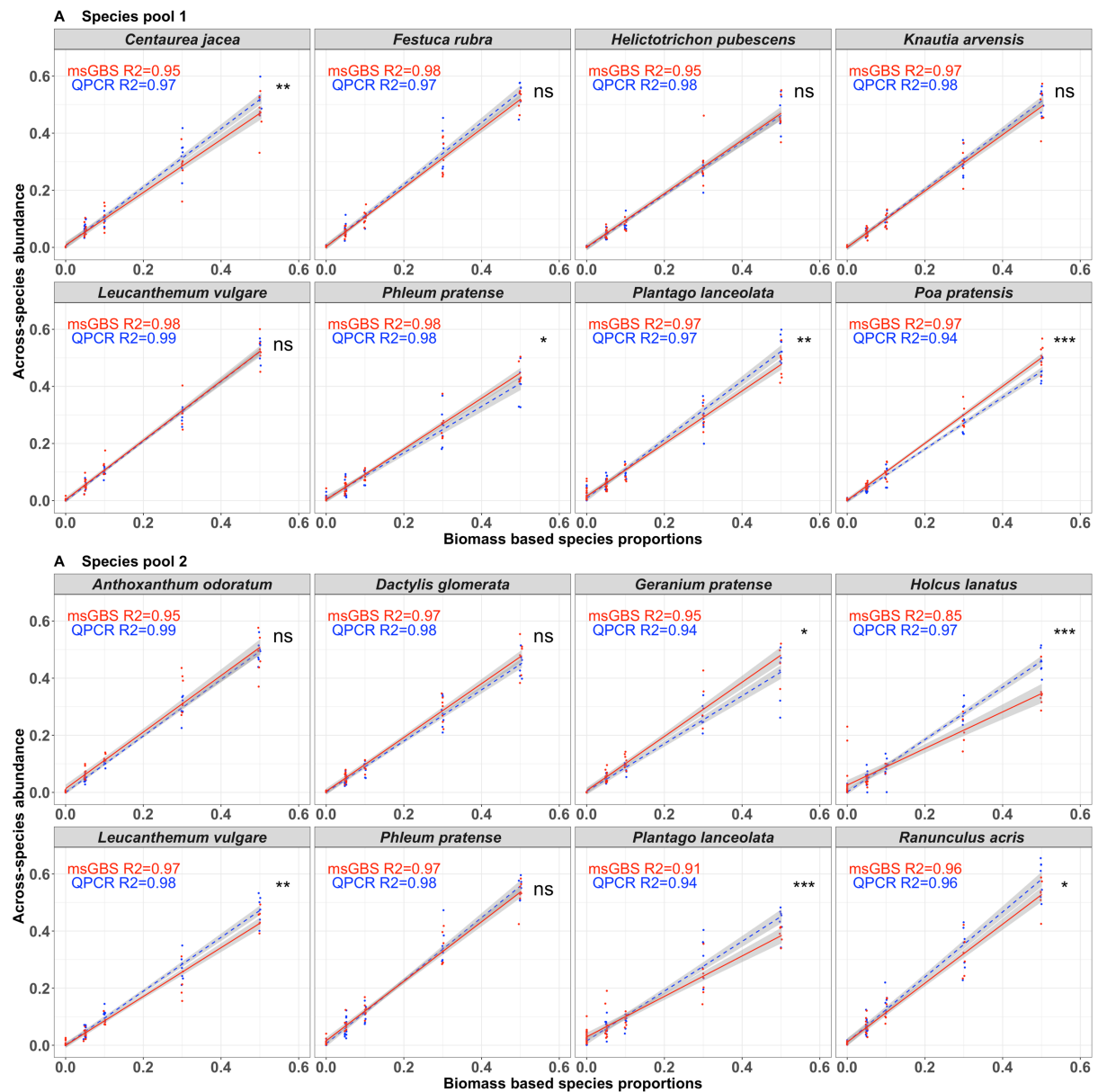875 as a table.
876

877

878
879 Figure 5 Comparison of the biomass-based species proportions to the msGBS (red, solid lines)
880 and qPCR (blue, dashed lines) Across-species abundance of the mock-mixture samples of
881 species pool 1 (A) 2 (B). The gray areas display the 0.95 confidence intervals. Asterisks note if
882 msGBS and qPCR regression models are significantly different (<0.001(***), 0.001-0.01(**)
883 and 0.01-0.1(*)).

884

	Congener group	Species present in mock sample	N	Average signal (%)	Species absent in mock sample	N	Average signal (%)	rFPS (%)
RevMet	*Papaver*	*Papaver somniferum*	6	4.07	*Papaver rhoeas*	6	0.07	1.64
	Ranunculus	*Ranunculus repens*	12	13.94	*Ranunculus acris*	12	4.97	35.62

	Congener group	Species present aboveground	N	Average signal	Species absent aboveground	N	Average signal	rFPS (%)
msGBS	*Centaurea*	*Centaurea jacea*	30	0.051220	*Centaurea scabiosa*	30	0.000049	0.10
	Cerastium	*Cerastium fontanum*	20	0.000100	*Cerastium arvense*	20	0.000005	2.91
		Cerastium glomeratum	4	0.000528				
	Cirsium	*Cirsium vulgare*	18	0.005523	*Cirsium arvense*	18	0.000802	14.53 *
	Convolvulus	*Convolvulus arvensis*	53	0.045996	*Convolvulus sepium*	53	0.000017	0.04
	Euphorbia	*Euphorbia esula*	8	0.000841	*Euphorbia helioscopia*	8	0.000003	0.41
	Galium	*Galium mollugo*	58	0.050836	*Galium aparine*	58	0.000037	0.07
	Geranium	*Geranium dissectum*	20	0.003020	*Geranium molle*	20	0.000007	0.24
					Geranium pratense	20	0.000025	0.84
	Ranunculus	*Ranunculus acris*	11	0.000336 **	*Ranunculus repens*	22	0.000144	43.96 ***
		Ranunculus bulbosus	18	0.000541 **				
	Rumex	*Rumex acetosa*	9	0.000309	*Rumex crispus*	9	0.000020	6.55
					Rumex thyrsiflorus	9	0.000096	31.12 ****
		Rumex acetosa	7	0.000710	*Rumex crispus*	8	0.000001	0.04
		Rumex thyrsiflorus	6	0.003044				
		Rumex acetosa	13	0.001469	*Rumex thyrsiflorus*	16	0.000348	15.99 ****
		Rumex crispus	9	0.003206				
	Trifolium	*Trifolium pratense*	15	0.004809	*Trifolium dubium*	15	0.000000	0.00
					Trifolium repens	15	0.000001	0.01
	Vicia	*Vicia cracca*	0	0.000000	*Vicia hirsuta*	24	0.000527	2.24
		Vicia sativa	24	0.023538	*Vicia sepium*	24	0.000007	0.03
		Vicia cracca	1	0.000019	*Vicia sepium*	15	0.000012	0.01
		Vicia sativa	14	0.053774				
		Vicia hirsuta	14	0.187578				

* high rFPS was caused by an isolated very high signal (within 1 of the 18 samples) suggesting that the 'absent' congener species *Cirsium arvense* was missed in the survey of this field plot or a seed accidently remained after washing and cleanup of this root sample.

** no interference between *R. acris* and *R. bulbosus*

*** high rFPS caused by the presence of *R. bulbosus*

**** interference between *R. acetosa* and R. *thyrsiflorus*

Table 3 Comparison of rFPS (%) obtained by RevMet (Peel et al., 2019) and non-calibrated msGBS within congener species groups. The average signal for RevMet is the bee pollen species abundance (%) and for msGBS it is the non-calibrated within-species root abundance of the *Dutch field study* root samples. N is the number of samples included in the comparison.