MOLECULAR ECOLOGY RESOURCES

Supplemental Information for:

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

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].

Addresses

[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, PObox 47, 6700 AA, Wageningen, The Netherlands
[3] Department of Systematic Botany and Functional Biodiversity, Institute of Biology, 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

MOLECULAR ECOLOGY
RESOURCES

Glossary

Clustering – The grouping of related sequences into a consensus sequence called clusters. A clustering algorithm is used to collapse DNA sequence reads within a certain percentage of similarity (in this supplement referred to as 95% identity) to create a consensus DNA sequence read called a (DNA) cluster.

Consensus sequence – A sequence in which SNP bases are represented as the most common one.

DNA cluster – a single DNA sequence that is the product the clustering of a collection of DNA sequence reads of high similarity.

SNP – Single Nucleotide Polymorphism.

Trimming – The removal of unwanted nucleotides from a sequence read. In our case we trimmed remaining adapter traces and low-quality nucleotides.

Sequence read mapping – Sequence read mapping is the process of comparing DNA sequence reads to a DNA reference based on a Burrows-Wheeler (Burrows & Wheeler, 1994) transformation aligner (like BWA, Bowtie2 and STAR).

qPCR – Quantitative Polymerase Chain Reaction is a common method mostly used to quantify gene expression but also commonly used for quantification of DNA. In general, it measures the amplification of a PCR product through time via a fluorescent signal from which the between samples relative DNA amounts are estimated

Universal primers – PCR primers that anneal to between species common DNA regions.

gDNA – genomic DNA

BLASTN search - Basic Local Alignment Search Tool is a is a program that uses an algorithm to searches a nucleotide database using a query DNA sequence read and identifies local similarity between sequences.

PCR duplicates – HTS only sequences a small proportion of the molecules in a (GBS) sequencing library. In msGBS, this sequencing library is the product of a PCR amplification step. When too many amplification cycles are used the HTS sequencing will sequence multiple copies of an original molecule. The library is over-amplified. PCR duplicates should be removed as it can cause a bias in the mapping dataset. Here we used 2x3 random nucleotides, called UMI's (Unique Molecule Identifier) in the adapter to identify and remove PCR duplicates. In our pipeline this process is done at a per sample and per cluster level.

Sequencing library – A collection of DNA fragments sequence ready for HTS sequencing. For msGBS Illumina Hiseq libraries this means gDNA inserts attached to sequence adapters that contain DNA anchors for annealing to the sequencing slide and primer binding sites for sequencing.

Library preparations and sequencing

All enzymatic reactions were performed in a PCR machine. During ligation a heated lid was used at lowest temperature available (30 degrees Celsius) and during restriction enzyme reactions a heated lid was used at 37 degrees Celsius. The msGBS libraries are constructed using two indexed adapters suitable for sequencing on a Illumina Hiseq platform (Fig S1 and Table S1). The adapters are designed according to van Gurp et al.(Van Gurp et al., 2016) with some modifications for enzyme choice and the use of 3N wobble nucleotides to identify PCR duplicates. To minimize the possibility of misidentification of samples, as a result of index sequencing- and synthesis errors, all pair-wise combinations of indices differed by a minimum of three mutational steps, index lengths were modulated from 4 to 6 bp to maximize nucleotide the balance of the first ten bases in order to improve the Illumina raw data analysis. As a consequence of using non-phosphorylated adapters, each DNA fragment–adapter connection contained a single stranded nick (Fig S1; yellow dot).

Extended lab protocol:

The protocol was altered in a number of ways of which the restriction enzymes used and the adapter design were the most relevant changes. Per sample, 300 ng gDNA was digested overnight (17hrs) at 37°C in a volume of 40 µL containing 1x FD buffer (B64, Thermo Scientific, The Netherlands), and 1 uL of both *Pac*I (TTAAT^TAA)(R0547L, New England Biolabs (NEB), England) and *Nsi*I (ATGCA^T)(R0127S, NEB). Following digestion, indexed "wobble" adapters were ligated to the fragments. Combining the 18 BA *Pac*I and 14 CO *Nsi*I adapters (Alpha DNA, Canada) resulted in a maximum of 252 index combinations (Fig. S1 and Table S1). The adapters each contain a 3bp random nucleotide region creating a small wobble after the adapter annealing process. For ligation 4 uL of both BA and CO indexed adapters (600 pg/uL), 6 uL T4 DNA ligase buffer, 0.5 uL T4 DNA ligase (M0202M, NEB) and 5.5 uL of distilled water was added to the 40 µL digestion mix. Ligation was performed for 3hrs at 22°C followed by 4°C overnight. All reactions were pooled, mixed and divided in 8 aliquots. The total volume of each pool was reduced to 40 µL using Qiaquick PCR cleanup (28104, Qiagen, The Netherlands) and size selected by a 0.8x Agencourt AMPureXP (A63880, Beckman coulter, Canada) purification using 22 µL lowTE buffer for elution. A nick translation reaction repaired the fragment–adapter nicks and re-assembled the 5-prime attached adapter strand which directly "unwobbles" the adapters using the opposite adapter strand as template (Fig S1). The enzymatic reaction (1h at 15°C) was performed in a volume of 25 µL containing 19.25 µL purified library, 2.5 µL 10 mM dNTP Mix (N0447L, NEB), 2.5 µL NEBuffer 2 and 0.75 µL DNA polymerase I (M0209, NEB). For each aliquot the library amplification was performed in four replicate 10 µL reactions containing 1 µL nick repaired DNA, 5 µL KAPA HiFi HotStart Sequence readyMix (KK2602, Roche Diagnostics, Switzerland) and 3 pmol of each Illumina PE PCR Primer.

Forward primer : 5'-aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatct-3'

Reverse primer : 5'-caagcagaagacggcatacgagatcggtctcggcattcctgctgaaccgctcttccgatct-3'
Temperature cycling consisted of 95°C for 3min followed by 14 cycles of 98°C for 10s, 65°C for 15s and 72°C for 15s with a final extension step at 72°C for 5min. Smaller DNA fragments are more efficiently amplified during PCR resulting in the sequencing of only a small portion of the genome. All replicate PCR products were pooled, concentrated using Qiaquick PCR cleanup, size selected by 0.8x Agencourt AMPureXP and quantified using a Qubit® dsDNA HS Assay Kit (Life technologies, USA). The size distribution and quality of the Libraries was assessed on a High Sensitivity DNA chip on a 2100 Bioanalyzer system (Agilent, USA). A qPCR quantification was performed by KAPA Library Quantification Kit for HTS (KK4844, KAPA Biosystems, USA) on a Biorad (The Netherlands) CFX96 Touch™ Real-Time PCR Detection System for optimal sequencing output. The libraries were spiked with 10% PhiX DNA to further increase the complexity of the libraries. 2x150bp Paired-End sequencing was executed by Novogene (Hongkong) on a Illumina (USA) Hiseq X-Ten sequencer.

qPCR *sensu Mommer et al. (2008) and Oram et al. (2018)*

From Oram et al. (2018):

In all samples, each species was separately amplified by real-time PCR with species-specific primer pairs (in triplicate). Primer pairs for *A. odoratum*, *F. rubra* and *L. vulgare* were used as described in Mommer et al. (2008). Primer pairs for *C. jacea, D. glomerata*, *G. pratense*, *H. lanatus*, *Haemanthus pubescens*, *K. arvensis*, *P. lanceolata*, *P. pratense*, *P. pratensis* and *R. acris* were developed using the same protocol as Mommer et al. (2008).

Real- time PCR reactions were performed with HOT FIREPol Eva Green (Solis BioDyne, Tartu, Estonia) qPCR Mix Plus with an addition of 0.94 µM MgCl2, a primer concentration of 60 nM for *A. odoratum* and *C. jacea* and 120 nM for all other species, and 4 ng genomic DNA for *P. lanceolata* or 1 ng genomic DNA for the other species, in a reaction volume of 20 µl. The qPCR program was as follows: 15 min at 95°C; then 41 cycles of 20 s at 95°C, 30 s at 62°C and 15 s at 72°C; and finally a melting curve analysis of 5 s per cycle, starting at 70°C and ending at 91°C with an increment of 0.5°C per cycle.

MOLECULAR ECOLOGY
RESOURCES

Species	Pool	Forward primers	Reverse primers
Anthoxanthum odoratum[*$]	2	5'-TCATGTACTGTTGTACTGCGAAG-3'	5'-GAATCAAGCTGGACAGTAAATGAC-3'
Avenula pubescens	1	5'-CTGGACGTTTCCCATGTTCT-3'	5'-GGTGGTACAGAGGTGGCAGT-3'
Centaurea jacea [$]	1	5'-CTCGCACATCCACGCACAC-3'	5'-TGCAGTGGTTTTCGTAGGAAGG-3'
Dactylis glomerata	2	5'-CAGGGCATTGAACTGATGATG-3'	5'-AGAAACTGGTGTGCGTCTGC-3'
Festuca rubra *[$]	1	5'-ACCGGAGATCGACAGCAAAACAG-3'	5'-TGTCCCTTGGTGGCGTTTTGG-3'
Geranium pratense	2	5'-ACCTTCGGGGAATCGTGTTA-3'	5'-TCGACCCAAGTGGTAAGGAG-3'
Holcus lanatus	2	5'-CAAGTTCGGAAGCCGTTAGG-3'	5'-GGACTCCAGTCCAGCGAAGT-3'
Knautia arvensis [$]	1	5'-GACCACAAAAGCAAGGAAGAA-3'	5'-CAAGGCAAGGAATCTCCAAG-3'
Leucanthemum vulgare *[$]	1,2	5'-AAACTCTACAGGCGTTCTTCC-3'	5'-ATTTCACTTCATAGCTCTTCACTG-3'
Phleum pratense [$]	1,2	5'-AGAGAGCAGGACACCGCCTA-3'	5'-GCCCTCTTGATTTTCGCATC-3'
Plantago lanceolata [$]	1,2	5'-GAGAAAGCAGTAGGAAACCACAGTG-3'	5'-GATCGAGATCTCTCACTCAAAACCC-3'
Poa pratensis [$]	1	5'-TGCACCCCTTCTGACTCTCA-3'	5'-GTGATAAGCGCGTCACGTTC-3'
Ranunculus acris	2	5'-CATTGCCACCTCTGCACTTC-3'	5'-TGAAACTTGCAGGTCCGAGA-3'

qPCR primers used for qPCR based quantification of across-species abundance; Table from Thesis Janneke Maria Ravenek (2015) Belowground species interactions and community effects in species-rich grasslands. PhD thesis, Radboud University Nijmegen, 220p. * Mommer et al. (2008) and used in Mommer et al. (2010). $ used in chapter 3 and 4 of thesis.

Extended bioinformatics

General information

Computations were executed on a local cluster node containing two Intel(R) Xeon(R) CPU E5-2450 0 @ 2.20GHz, 512GB RAM and 14 TB hard disk space, using ubuntu 16.04. Editing scripts and debugging was locally performed using PyCharm Professional 2017 2.2 using Python (Python core Team, 2015). R (Suhl et al., 2014) was executed using Rstudio (RStudio Team, 2016).

The msGBS data processing can be partitioned in 5 processes which are outlined in figure 2. Process 1 describe the pre-processing of the raw sequence reads. Process 2 describe the creation and BLASTN filtering of the meta-reference genome from monoculture root material. Process 3 is the mapping of all sequence reads (Fig. 2, product 1) to the filtered meta-reference (Fig. 2, product 2) which result in the BAM alignment file (Fig. 2, product 3). Process 4 is the processing of the UMI's in the BAM alignment header line, the conversion of the BAM file to CSV format and the monoculture-based cluster filtering which identifies and removes, between species, homologous clusters within the CSV FILE. In the final process 5 the within-species abundance is calculated (Fig. 2, product 5a) and optionally calibrated using the 'calibration key' which results in across-species abundance (Fig. 2, product 5b).

The analysis in this article were performed using a series of scripts and commands as described in detail below and are available on https://github.com/NielsWagemaker/scripts_msGBS branch msGBS-1.0. A new, more efficient and easy to install (Anaconda, 2016), snakemake (Köster et al., 2012) version of the pipeline is also made available on GitHub (https://github.com/NielsWagemaker/scripts_msGBS using the msGBS-snake branch).

The msGBS-snake Bioinformatics pipeline, essentially doing the same as the msGBS-1.0 scripts, can be installed in four easy steps. Example files and further instructions can be found on https://github.com/NielsWagemaker/scripts_msGBS master branch (msGBS-snake branch). This whole msGBS-snake pipeline consists of several independent commands and scripts which are executed using the snakemake (Köster et al., 2012) workflow management system. The combined use of conda and snakemake makes setting up and running the pipeline accessible for beginning Linux users:

- Step one is to clone the msGBS github repository from your linux based terminal.

```
git clone https://github.com/NielsWagemaker/scripts_msGBS.git
```

- Step two is to install conda (Anaconda, 2016) on your linux system.
    - See for instructions: https://docs.conda.io/projects/conda/en/latest/user-guide/install/linux.html
- Step three is to create a environment in which all dependencies will be installed and from which the pipeline is run and subsequent activation of the environment
    - conda env create -f src/env/environment.yaml --name msGBS

  - conda activate msGBS
- Step four is Run the pipeline
  - Make barcode file
  - Fill in the config.yaml file
  - Run <snakemake -j 12> (and will use 12 cores)

The Analysis steps and scripts of the original msGBS-1.0 pipeline as used for this article are explained in detail below:

Process 1 : Sequence read pre-processing

Demultiplexing

The first step of the pre-processing (Fig. 2, process 1) was performed by the script Demultiplex_msGBS.py. The input files of the script are the zipped fastq sequence read1 (R1) and sequence read2 (R2) files obtained from Novogene and a tab delimited 'index or barcode sheet' that contains sample information (a.k.a. sequence read group information), the indices for the adapters (BA and CO), the number of UMI nucleotides for identification of PCR duplicates (BA and CO) and the used enzymes (BA and CO). An example barcode sheet along with instructions on how to create the correct file type can be found on GitHub. The script removes the index and UMI nucleotides from the R1 and R2 sequence reads and adds them to the sequence information line of each sequence read (BC:Z: and RN:Z: info tags, respectively). The index information is used to extract the sequence read group information from the index file (RG:Z: info tag). The output files of the script are the zipped fastq sequence read1 (R1) and sequence read2 (R2) demultiplexed output files.

Example raw data:

read1:

```
@ST-E00317:403:H53KHCCXY:6:1101:1966:1309 1:N:0:NCCCCCCC
NCCCTAATCATTCCTTTCTTTTCGTTAGCGTGTCATATCACTTTTTTAATTTTAGTTTTTCATTTGTTTTGCATCTTTTCAAATAGTGTTTCGTACACTGAATTAGTTTTTTGAATTTTTTAAAATGTCCAAAAATTATTTTTCTCGAAA
+
#AAAFJJFJJAFAA<<FJFJJFJJFJ-FJFFA<FF<J-F-FFJJ<FFJ-<A<F-AAF-F<FFJFJF-AJJJJ<AJAA-F7A7-<-F-7-AA<-7AF-77--7FAF-AFFA--7-7---FJJA--AJF--7<AFAJF-AJAFFAFA-----
```

read2:

```
@ST-E00317:403:H53KHCCXY:6:1101:1966:1309 2:N:0:NCCCCCCC
NTCCAGCCTGCATCTCATTTTTCTTGCAAAAAACTAAAATGATTTTCCTGGAATTTTTCGAGAAACATTATTTTTGGACATTTCAAAAAATCCTAAAAACTAATTCAGTGTACGAAACACTATTTGAAAAGCTGCAAAACAAATGAAAAA
+
#AAAAJFJJJJJ<FAJFFJ<A-FJ-FJJJFFJF-<JFJ77FF--<-FJ<-7AFJ-7-7<7AA-F<-<AFJ<7A-A<JFF-AFF-<F-AFA---7FJA7JA-F--<FJF--AJFJA-AJFJ<A<-AAFA-7--<777A--F7F7F--A-<<
```

Example demultiplex data:

read1:

```
@ST-E00317:403:H53KHCCXY:6:1101:1966:1309       BC:Z:CTAAT  BC:Z:CAGC   RG:Z:H53KHCCXY_6_ratio_45_jena2016  ST:Z:Crick  RN:Z:NCC_NTC
ATTCCTTTCTTTTCGTTAGCGTGTCATATCACTTTTTTAATTTTAGTTTTTCATTTGTTTTGCATCTTTTCAAATAGTGTTTCGTACACTGAATTAGTTTTTTGAATTTTTTAAAATGTCCAAAAATTATTTTTCTCGAAA
+
JAFAA<<FJFJJFJJFJ-FJFFA<FF<J-F-FFJJ<FFJ-<A<F-AAF-F<FFJFJF-AJJJJ<AJAA-F7A7-<-F-7-AA<-7AF-77--7FAF-AFFA--7-7---FJJA--AJF--7<AFAJF-AJAFFAFA-----
```

read2:

```
@ST-E00317:403:H53KHCCXY:6:1101:1966:1309       BL:Z:CTAAT  BR:Z:CAGC   RG:Z:H53KHCCXY_6_ratio_45_jena2016  ST:Z:Crick  RN:Z:NCC_NTC
TGCATCTCATTTTTCTTGCAAAAAACTAAAATGATTTTCCTGGAATTTTTCGAGAAACATTATTTTTGGACATTTCAAAAAATCCTAAAAACTAATTCAGTGTACGAAACACTATTTGAAAAGCTGCAAAACAAATGAAAAA
+
JJJJ<FAJFFJ<A-FJ-FJJJFFJF-<JFJ77FF--<-FJ<-7AFJ-7-7<7AA-F<-<AFJ<7A-A<JFF-AFF-<F-AFA---7FJA7JA-F--<FJF--AJFJA-AJFJ<A<-AAFA-7--<777A--F7F7F--A-<<
```

Not in article referenced figure. The example raw sequence read 1 and 2 show the sequence read information line, the sequence read sequence line, a separation line (+), and the sequence read quality line. The sequence read information line is the combination of the Sequencer code (@ST-E00317:403:), the Flowcell code (H53KHCCXY), lane number (6), sequence read coordinates (1101:1966:1309) and Illumina index information (1:N:0:NCCCCCCC, which we not use as we use 'inline indices'), respectively. The example demultiplex data shows the additional information tags; BA and CO indices (BC:Z:CTAAT BC:Z:CAGC), sequence read group information (a.k.a. sample name)(RG:Z:H53KHCCXY_6_ratio_45_jena2016), strand identifier (ST:Z:Crick) and wobble nucleotides (RN:Z:NCC_NTC).

Adapter trimming

The second step of the pre-processing is the removal of adapter remnant sequences and the trimming of low quality nucleotides (Fig. 2, process 1). The program AdapterRemoval (Schubert, Lindgreen, & Orlando, 2016) was used following the below mentioned parameters.

```
AdapterRemoval --file1 R1_demultiplex_NGmerge_H53KHCCXY_s_6_fastq.txt.gz --file2 R2_demultiplex_NGmerge_H53KHCCXY_s_6_fastq.txt.gz --basename R1_N12_ H53KHCCXY _NG2_no_adapter_ --trimns --trimqualities --minquality 10 --minlength 100 --adapter-list adapters.txt --gzip
```

Process 2 : Meta-reference creation and BLASTN filtering

Meta-reference creation

For each monoculture sample in the experiment the python script Make_reference_msGBS.py assembles clusters and creates a reference output file (Fig. 2, process 2). The clusters of all monoculture references are combined into a single meta-reference file while retaining original monoculture identifier names and cluster numbers.

The script first finishes the pre-processing using NGmerge (Gaspar, 2018) in 'stitch' mode for merging the R1 and R2 sequence reads using the default parameters (Fig. 2, process 1). In default NGmerge requires that a valid alignment to have a minimum overlap of 20 bp and a maximum of 10% mismatches in the overlap region (-m 20 -p 0.1). NGmerge outputs the merged sequence reads and the R1 and R2 non-merged sequence reads. A log file (in our case a 190 GB file with our 16GB merged.gz file) can optionally be produced for evaluation of all merging events. The quality lines of the reads are retained during this process. In the overlapping stretches the highest quality bases are retained.

The non-merged sequence reads were then joined and a 'NNNNNNNN' sequence was inserted between the R1 and R2 sequence reads. The R2 sequence reads were not reverse complemented. The combined set of merged and joined sequence reads of all samples is product

1 of figure 2. For each monoculture sample the merged- and joined sequence reads were extracted from Figure 2, product 1 and separated into single files per monoculture and dereplicated using Vsearch (Rognes, Flouri, Nichols, Quince, & Mahé, 2016) using the following parameters : `–derep_fulllength <file_fa.gz_in> --sizeout --minuniquesize 2 –output <output_fa_file>`. The resulting merged- and joined files were then combined into a single file per monoculture. These files were then sorted with Vsearch using the following parameters: `vsearch --sortbylength <input_fa_file> --output <output_fa_file>.` Clustering with 95% identity was performed using `vsearch –cluster_smallmem <input_fa_file> –id 0.95 --centroids <output_fa_file> –sizeout --strand both`. Because of the low sequencing input for the Jena experiment we kept the minunique size at 2 to prevent even lower number of clusters generated. This resulted in, on average, one cluster per 12 input reads. For the Dutch experiment we used minunique size 3. This resulted in, on average, one cluster per 151 input reads. Whether this dramatic effect in number of clusters generated is mainly caused by the adjustment of this parameter or by the different origin of the monoculture material (root versus leaf, respectively) is unknown.

The resulting file was then transformed by the custom python script rename_fast.py. Finally, the references of all 'monocultures' were combined into a single unfiltered meta-reference file named ref.fa. This meta-reference file is indexed using Samtools (Li et al., 2009) by using the following parameters : `samtools faidx <input_fa_file>` creating a ref.fa.fai index file.

BLASTN filtering

The meta-reference filtering step (Fig 2, process 2) is performing by a local BLASTN search using the meta-reference against the NCBI nr database using the following command: `blastn –query <path_to>ref.fa –db nt –out <path_to>outputblast_kingdoms.txt –num_alignments 1 –num_thsequence reads 12 –outfmt '6 qseqid sseqid pident evalue bitscore sskingdom sscinames length sstart send '`. The BLAST output file contains only the highest hit and some BLAST characteristics; Subject Scientific Kingdom and the Subject Scientific Name.

A custom script named blastN_parse_ref_msGBS.py was used to parse the output file and filter out all clusters that were annotated as Bacteria, Archaea or Viruses. For Eukaryota the genus names in the BLAST output file were used to identify arbuscular mycorrhizal (AM) Fungi (based on an AM Fungi genera list in the script), other Fungi (based on the genera extracted from http://www.mycobank.org/), and plants (based on a genera list extracted from http://www.theplantlist.org). A minimum alignment length of 40 bp and an 'expect value' of smaller than 1e-20 of the BLASTN hit was required to accept the annotation. A .txt file was written for each of the above groups containing the clusters that were filtered out the reference. The

script also creates text files containing a list of the genus names. Non-filtered identified clusters and clusters that gave no hit to the NR database were retained in the meta-reference (Fig. 2, product 2).

Process 3 : Mapping

A custom python script named Map_STAR_msGBS.py was used for mapping of the pre-processed reads (Fig. 2, product 1) to the meta-reference file ref.fa (Fig. 2, product 2). Map_STAR_msGBS.py uses the R script STAR (Zaleski et al., 2012). STAR creates a mapping index genome for merged- and joined sequence reads separately using the following parameters:

```
STAR --runThsequence readN 6 --runMode genomeGenerate --genomeDir <genome_directory> --genomeFastaFiles <genome_directory/fasta_file> --genomeSAindexNbases <locally calculated> --genomeChrBinbits <locally calculated>.
```

The sequence read mapping by STAR for joined and merged sequence reads independently is performed using the following parameters:

```
STAR --runThsequence readN 16 --genomeDir <genome_directory> --sequence readFilesIn <non-assembled R1 sequence reads> or <merged sequence reads> --outSAMattributes NM MD AS --outSAMtype SAM --outFileNamePrefix <prefix of output files> --outFilterMatchNminOverLsequence read 0.95 --clip3pNbases 1 --outSAMorder PairedKeepInputOrder --outFilterMultimapScoreRange 0 --alignEndsType EndToEnd --scoreGapNoncan 0 --scoreGapGCAG 0 --scoreGapATAC 0 --scoreDelOpen 0 --scoreDelBase 0 --scoreInsOpen 0 --scoreInsBase 0 --alignMatesGapMax 20.
```

Using these parameters allows the multimapping of a read to a maximum of 10 clusters but only when they have identical high alignment scores. These setting allow the monoculture-based cluster filtering to be possible. Otherwise detection of between monoculture homologous clusters would not be possible.

After mapping and file reformatting the Map_STAR_msGBS.py script sorts the SAM file using Sambamba (Tarasov, Vilella, Cuppen, Nijman, & Prins, 2017) and saves the output as BAM file (Fig 2. product 3).

Process 4 : Post-processing of read mapping data

The marking of PCR duplicate sequence reads

The custom script Mark_PCR_duplicates.py uses the UMI nucleotides in de RN:Z: tag of the demultiplexed sequence read pairs to identify PCR duplicates (Fig 1, step 7). When both R1 and R2 wobble code are identical between individual sequence read-pairs these are considered PCR duplicates and marked as 'is_duplicate' in the BAM file. Further when sequence reads AS (Alignment Score) was below 0.8 x cluster length the sequence read was marked as 'qc_fail' in BAM file.

BAM to CSV conversion

The custom script msGBS_STATS.py parses the BAM file and exports a .CSV output file containing the mapped sequence reads counts per clusters for all samples in the experiment. The script discards sequence reads that are marked 'qc_fail' and 'is_duplicate' and does not extract clusters to which less then 10 reads over *all* samples mapped.

Monoculture-based cluster filtering

The monoculture-based cluster filtering, the final step of Figure 2 process 4, is performed by the custom script Parse_csv.py which also produced the products of process 5. The complete synthax is `parse_csv.py -i /Users/NielsWagemaker/Dropbox/Documents/PROJECTEN/HANS/SGBS_barcoding/article_2018/DATA/mapping_Feb_2019_ref095_map095/out_dedup08_stats_726605_clusters.csv -op <PATH>\msGBS_run1_ -os f8_f15_1000_pool1_no1_STD.csv -f1 8 -f2 15 -f3 1000 -p 1 -e EXTRA` which includes input file, output suffix and prefix, the monoculture-based cluster filtering parameters f1, f2 and f3, the pool for process 5 analysis and -e for optional extra output. The monoculture cluster filtering is independent of species pool as all monocultures are evaluated simultaneously.

Process 5 : Non-calibrated and calibrated analysis

After monoculture-based cluster filtering `parse_csv.py` renames the 'jenamono1-13' samples to actual species names. The within-species abundances of the mock mixture root samples are calculated and saved in e10_tussen_sum_REL_p<pool_f1_f2>.csv. The calibration key is calculated from the calibration sample read mapping counts (from filtered CSV file) and subsequently used to do calibrated analysis. The across-species abundances of the mock mixture root samples are saved in <prefix>FINAL<suffix>.

MOLECULAR ECOLOGY
RESOURCES

MOLECULAR ECOLOGY
RESOURCES

Figures and Tables

Library prep and adapters

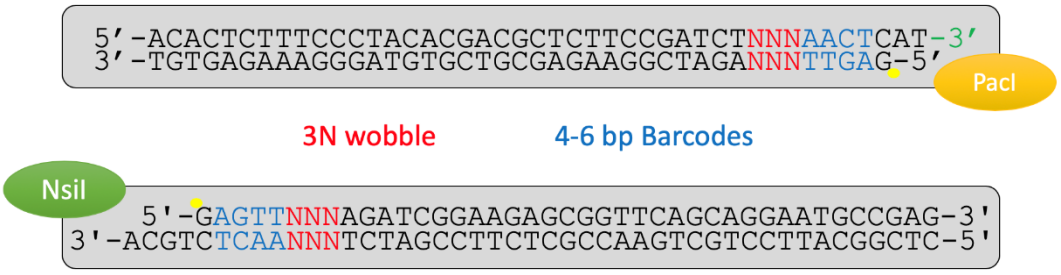


Figure S1. Overview of the BA (upper) and CO (lower) indexing adapters. The blue nucleotides represent the variable index nucleotides used for demultiplexing (Table S1). The red nucleotides represent the random unique molecule identifier (UMI) nucleotides used for PCR duplicate identification. Before nick repair these random nucleotides cause a local mismatch between the adapter strands. The *Pac*I (yellow oval) and *Nsi*I (green oval) represent the complementary overhang of the adapters with the respective restriction enzymes used in msGBS. The yellow dots represent nick's in the dsDNA before nick repair.

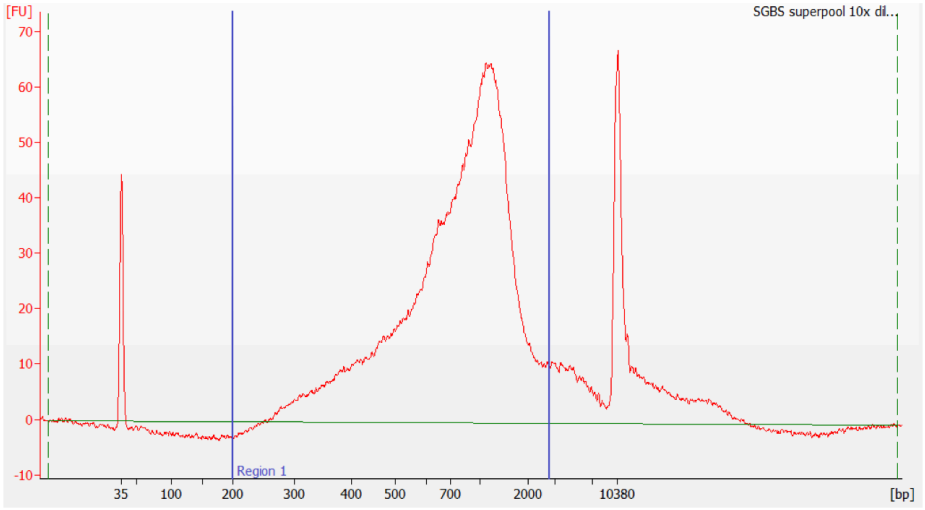


Figure S2. Bioanalyzer result of *the Jena study* msGBS pooled library. Internal Bioanalyzer markers are visible at 35bp and 10,380bp. Smallest library fragments are around 200bp, no adapter dimer was detected (146bp). Above 2,000bp non-amplified fragments are visible. X-axis displays fragment length in bp, y-axis displays fluorescent signal units (FU).

MOLECULAR ECOLOGY
RESOURCES

Meta-reference creation and over-all read mapping statistics

1	*Achillea millefolium*	26	*Cerastium fontanum*	51	*Geranium pratense*	76	*Phleum pratense*	101	*Tanacetum vulgare*
2	*Agrimonia eupatoria*	27	*Cerastium glomeratum*	52	*Glechoma hederacea*	77	*Picris hieracioides*	102	*Taraxacum sectie ruderalia*
3	*Agrostis capillaris*	28	*Cichorium intybus*	53	*Helictotrichon pubescens*	78	*Pimpinella major*	103	*Tragopogon pratensis s. orientalis*
4	*Ajuga reptans*	29	*Cirsium arvense*	54	*Heracleum sphondylium*	79	*Plantago lanceolata*	104	*Trifolium dubium*
5	*Allium vineale*	30	*Cirsium vulgare*	55	*Holcus lanatus*	80	*Poa pratensis*	105	*Trifolium pratense*
6	*Alopecurus pratensis*	31	*Convolvulus arvensis*	56	*Isatis tinctoria*	81	*Poa trivialis*	106	*Trifolium repens*
7	*Anisantha sterilis*	32	*Convolvulus sepium*	57	*Jacobaea vulgaris s. vulgaris*	82	*Potentilla reptans*	107	*Trisetum flavescens*
8	*Anthoxanthum odoratum*	33	*Corylus avellana*	58	*Knautia arvensis*	83	*Ranunculus acris*	108	*Urtica dioica*
9	*Anthriscus sylvestris*	34	*Crataegus monogyna*	59	*Lamium album*	84	*Ranunculus bulbosus*	109	*Valerianella carinata*
10	*Arenaria serpyllifolia*	35	*Crepis biennis*	60	*Lamium maculatum*	85	*Ranunculus repens*	110	*Valerianella locusta*
11	*Arrhenatherum elatius*	36	*Cruciata laevipes*	61	*Lamium purpureum*	86	*Rhinanthus alectorolophus*	111	*Verbascum nigrum*
12	*Bellis perennis*	37	*Cynosurus cristatus*	62	*lathyrus pratensis*	87	*Rhinanthus minor*	112	*Veronica arvensis*
13	*Brassica nigra*	38	*Dactylis glomerata*	63	*leucanthemum vulgare*	88	*Rosa canina*	113	*Veronica chamaedrys*
14	*Bromopsis inermis s. inermis*	39	*Elymus repens*	64	*lolium perenne*	89	*Rubus caesius*	114	*Veronica hederifolia*
15	*Bromus hordeaceus s. hordeaceus*	40	*Equisetum arvense*	65	*lotus corniculatus*	90	*Rumex acetosa*	115	*Veronica persica*
16	*Calamagrostis epigejos*	41	*Euphorbia esula*	66	*lysimachia nummularia*	91	*Rumex crispus*	116	*Vicia cracca*
17	*Campanula rapunculus*	42	*Euphorbia helioscopia*	67	*Malva sylvestris*	92	*Rumex thyrsiflorus*	117	*Vicia hirsuta*
18	*Capsella bursa-pastoris*	43	*Festuca arundinacea*	68	*Medicago lupulina*	93	*Salvia pratensis*	118	*Vicia sativa s. nigra + s. segetalis*
19	*Cardamine hirsuta*	44	*Festuca pratensis*	69	*Mercurialis annua*	94	*Sanguisorba minor*	119	*Vicia sepium*
20	*Cardamine pratensis*	45	*Festuca rubra + f. arenaria*	70	*Myosotis arvensis*	95	*Securigera varia*		
21	*Carduus crispus*	46	*Ficaria verna s. Verna*	71	*Origanum vulgare*	96	*Senecio vulgaris*		
22	*Carex hirta*	47	*Galium aparine*	72	*Papaver dubium*	97	*Sonchus asper*		
23	*Centaurea jacea*	48	*Galium mollugo*	73	*Pastinaca sativa s. sativa*	98	*Sonchus oleraceus*		
24	*Centaurea scabiosa*	49	*Geranium dissectum*	74	*Persicaria amphibia*	99	*Stellaria media*		
25	*Cerastium arvense*	50	*Geranium molle*	75	*Phalaris arundinacea*	100	*Symphytum officinale*		

Table S1 Species list of the *Dutch field study* reference species. Coloured are the congener species groups; for 11 congener groups root mixture samples were available for within congener msGBS signal comparison.

	Index	
	BA - *PacI*	**CO - *NsiI***
1	AACT	AACT
2	CCAG	CCAG
3	TGGC	TTGA
4	GAAC	GGTC
5	GCGT	ACTA
6	ATCAC	CAGC
7	CTAAT	TGAT
8	TGCGG	GTCG
9	AAGAT	ATAC
10	CCACC	CGCA
11	GTAGG	TATG
12	TGTACC	GCGT
13	AATCGG	GTAC
14	ACTGCC	AGCT
15	CTCAGG	
16	TATGCG	
17	GTGAGC	
18	CGTCAC	

Table S2. Adapter indices of the BA-*Pac*I and CO-*Nsi*I adapters. A maximum number of combinations of 18*14 = 252 can be made. And thus, a maximum of in total 252 samples can be processed in a single Sequencing lane.

MOLECULAR ECOLOGY
RESOURCES

		Summarized field survay				
		location 1	location 2	location 3	location 4	location 5
		ND115	ND102	ND56	ND46	EN97
		5 plots	5 plots	5 plots	5 plots	5 plots
	Average number of species present	24.8	28.2	22.1	22.4	23.8
Genera group	Species in meta-reference	Species present / absent				
Centaurea	*Centaurea jacea*	■	■	■	■	■
	Centaurea scabiosa	□	□	■	■	■
Cerastium	*Cerastium arvense*	■	■	□	□	□
	Cerastium fontanum	■	■	■	■	■
	Cerastium glomeratum	■	■	■	■	■
Cirsium	*Cirsium-arvense*	□	□	■	■	■
	Cirsium vulgare	■	■	■	■	■
Convolvulus	*Convolvulus arvensis*	■	■	■	■	■
	Convolvulus sepium	■	□	□	□	□
Euphorbia	*Euphorbia esula*	■	■	■	■	■
	Euphorbia helioscopia	□	□	■	■	■
Galium	*Galium aparine*	□	□	■	□	□
	Galium mollugo	■	■	■	■	■
Geranium	*Geranium dissectum*	■	■	■	■	■
	Geranium molle	□	□	■	□	■
	Geranium pratense	□	□	■	□	■
Ranunculus	*Ranunculus acris*	■	■	■	■	■
	Ranunculus bulbosus	■	■	■	■	■
	Ranunculus repens	■	■	□	□	■
Rumex	*Rumex acetosa*	■	■	■	■	■
	Rumex crispus	□	□	■	■	■
	Rumex thyrsiflorus	□	■	■	□	□
Trifolium	*Trifolium dubium*	■	■	■	□	■
	Trifolium pratense	□	■	■	■	■
	Trifolium repens	□	□	□	□	■
Vicia	*Vicia cracca*	■	■	■	■	■
	Vicia hirsuta	■	■	□	■	□
	Vicia sativa	■	■	■	■	■
	Vicia sepium	■	■	□	□	□

■ Species present at location
□ Species absent at location

Table S3 The congener species of 11 congener groups and their aboveground presence at location 1 to 5. The total species present in the 5 subplots of each location combined ranged from 22.1 to 24.8 species.

MOLECULAR ECOLOGY RESOURCES

			Sequenced	Assembled			Mapped (BAM)	
Hiseq sequence run		# samples	Total reads	Reads	Total Reads	(%)	Reads	(%)
Jena	Monoculture samples	13	217,171,278	15,534,803	181,555,188	84%	112,722,150	62%
	Mock samples	111		145,748,073				
	Calibration samples	20		20,272,312				
Dutch - run 1	Monoculture samples	119	378,265,715		309,202,157	82%	184,373,290	60%
Dutch - run 2	Field mixture root samples	105	291,588,907		218,351,794	75%	70,334,580	32%

			Extracted (CSV)		After filtering (CSV)		
Hiseq sequence run		# samples	Total reads	Average reads	Total reads	(%) of assembled	Average reads
Jena	Monoculture samples	13	9,446,804	726,677	3,275,252	21%	251,942
	Mock samples	111	62,913,352	566,787	19,780,209	14%	178,200
	Calibration samples	20	14,083,211	704,161	2,677,063	13%	133,853
Dutch - run 1	Monoculture samples	119	178,922,440	1,503,550	54,702,836	18%	459,688
Dutch - run 2	Field mixture root samples	105	68,255,194	650,049	12,772,460	6%	121,642

Table S4 Overview of the reads statistics per study and sample type. Displayed are the number of total raw PE sequence reads, assembled reads, mapped reads (in BAM file) and extracted read counts (in CSV file). Mapped sequence reads were not extracted from the BAM file to CSV file when they were annotated as is_duplicate / is_qual_fail or when the cluster to which the sequence reads mapped did not meet the minimum of 10 mapped sequence reads over all samples.

			Sequence reads										
Species	Pool		input			after dereplication			Length			Dereplication	Clusters
	1	2	merged	joined	combined	merged	joined	combined	min	max	average	factor	
Plantago lanceolata	X	X	404,729	530,158	934,887	77,090	59,590	136,680	95	290	232	6.8	99,799
Ranunculus acris		X	513,705	882,415	1,396,120	87,675	103,841	191,516	98	290	246	7.3	120,982
Knautia arvensis	X		496,177	813,154	1,309,331	89,902	101,587	191,489	98	290	244	6.8	122,643
Geranium pratense		X	29,374	62,133	91,507	5,401	4,225	9,626	103	290	233	9.5	8,475
Centaurea jacea	X		312,736	500,444	813,180	61,095	50,691	111,786	99	290	235	7.3	88,841
Dactylis glomerata		X	412,382	693,101	1,105,483	73,565	65,308	138,873	94	290	236	8.0	88,045
Anthoxanthum odoratum		X	285,948	572,313	858,261	52,879	56,289	109,168	98	290	241	7.9	77,890
Holcus lanatus		X	531,214	1,007,155	1,538,369	62,746	85,250	147,996	97	290	248	10.4	108,172
Festuca rubra	X		212,833	400,935	613,768	40,546	34,272	74,818	100	290	234	8.2	54,025
Helictotrichon pubescens	X		371,663	508,258	879,921	65,381	43,373	108,754	101	290	230	8.1	73,419
Poa pratensis	X		1,013,253	1,515,473	2,528,726	137,900	124,900	262,800	98	290	237	9.6	169,324
Leucanthemum vulgare	X	X	218,333	389,304	607,637	40,529	33,225	73,754	99	290	237	8.2	55,516
Phleum pratense	X	X	1,144,362	1,713,251	2,857,613	210,085	172,209	382,294	94	290	235	7.5	260,885
			Total reads		15,534,803	Total reads		1,939,554	Total clusters in meta reference				1,328,016

Table S5. *The Jena study* meta-reference input and output reads (merged- and joined) for all species in the meta-reference before and after dereplication. The final column shows the number of Clusters generated after clustering with 95% identity using Vsearch. *Geranium* had low sequence read input which resulted in subsequent low clusters numbers.

MOLECULAR ECOLOGY RESOURCES

BLASTN filtering

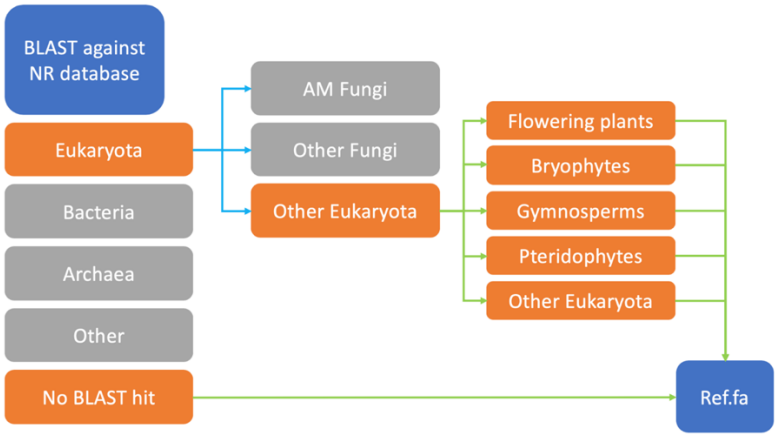


Figure S3. The clusters of the meta-reference were filtered based on the outcome of a BLASTN search against the NCBI NR database. A custom script named blastN_parse_ref_msGBS.py was used to parse 'Kingdom' information in the BLASTN output file and filter out all clusters that were annotated as Bacteria, Archaea, Viruses and Phages. For Eukaryota the genus names were used to further identify AM Fungi and other Fungi (genus names retrieved from http://www.mycobank.org/) and plants (genus names retrieved from http://www.theplantlist.org). Orange boxes represent clusters that are retained in the metareference, the gray boxes represent clusters that are filtered out.

SPECIES	Pool		Metareference clusters				Removed clusters				
	1	2	Before BLASTN	After BLASTN	Removed	Removed (%)	Bacteria	Archaea	Viruses	AM Fungi	other Fungi
Plantago lanceolata	X	X	99,799	96,210	3,589	3.6%	982	4	3	2,523	77
Ranunculus acris		X	120,982	118,890	2,092	1.7%	260	0	0	1,802	30
Knautia arvensis	X		122,643	120,533	2,110	1.7%	385	0	6	1,660	59
Geranium pratense		X	8,475	8,413	62	0.7%	27	0	0	30	5
Centaurea jacea	X		88,841	87,462	1,379	1.6%	501	2	3	832	41
Dactylis glomerata		X	88,045	87,528	517	0.6%	244	1	4	172	96
Anthoxanthum odoratum		X	77,890	77,519	371	0.5%	248	0	0	52	71
Holcus lanatus		X	108,172	107,697	475	0.4%	191	0	1	214	69
Festuca rubra	X		54,025	53,630	395	0.7%	207	0	27	102	59
Helictotrichon pubescens	X		73,419	72,819	600	0.8%	240	2	0	310	48
Poa pratensis	X		169,324	167,898	1,426	0.8%	626	0	4	603	193
Leucanthemum vulgare	X	X	55,516	55,226	290	0.5%	80	1	1	198	10
Phleum pratense	X	X	260,885	258,929	1,956	0.7%	1,127	2	10	594	223
			1,328,016	1,312,754	15,262	1.1%	5,118	12	59	9,092	981
							34%	0%	0%	60%	6%

Table S6. Number of *the Jena study* meta-reference clusters before and after BLASTN (against NR database) filtering including the 'annotation' of the BLASTN hit per monoculture. Most removed clusters were annotated as AM Fungi with Bacteria as second. Almost all AM Fungi clusters that were removed had hits to *Rhizophagus irregularis* strain DAOM_181602=DAOM_197198 genome.

Mapping

None

Monoculture-based cluster filtering

In total the *Monoculture-based* cluster filtering of the *the Jena study* meta-reference removed 513,238 of the 726,605 clusters in the CSV file resulting in the filtered CSV file (Fig 2, product 4). 88,217 clusters were removed due to the prefilter, 389,679 by application of filter parameter f1 and 35,342 after evaluation of filter parameter f2 (Table S6). The filter parameter f2 is only effective for clusters to which enough reads are mapped; we therefor set filter parameter f1 to 8 reads; a balance between effective monoculture-based cluster filtering and the yield of high enough read counts for final quantification. Still, if no non-target reads are recorded it can be that they is not detected due to the low read count. As discussed the low number of *Geranium* reads is expected to result in higher FPS; the *Geranium* root material will contain *Geranium* reads that homologous to other species clusters which were not identified due to the low monoculture read numbers.

MOLECULAR ECOLOGY
RESOURCES

A

After process	Product	After sub-process	Discarded clusters	Clusters in metareference	Clusters in CSV
Reference creation				1,328,016	
BLASTN filtering	2		15,262	1,312,754	
Extract mapping data			586,149		726,605
Monoculture based cluster filtering		Prefilter	88,217		638,388
		Parameter f1=8	389,679		248,709
		Parameter f2=15	35,342		213,367
	4	Parameter f3=1000			213,367

B

Species	Pool 1	Pool 2	Clusters in CSV file after monoculture based cluster filtering
Plantago lanceolata	X	X	6,599
Ranunculus acris		X	24,842
Knautia arvensis	X		23,940
Geranium pratense		X	117
Centaurea jacea	X		8,154
Dactylis glomerata		X	13,548
Anthoxanthum odoratum		X	8,047
Holcus lanatus		X	43,629
Festuca rubra	X		5,836
Helictotrichon pubescens	X		5,111
Poa pratensis	X		34,297
Leucanthemum vulgare	X	X	2,957
Phleum pratense	X	X	36,290
			213,367

Table S7 Number of clusters of *the Jena study* in the meta-reference before and after BLASTN filtering and in the extracted CSV file before and after monoculture-based cluster filtering (A). The number of clusters in the meta-reference after monoculture-based cluster filtering, split per monoculture.

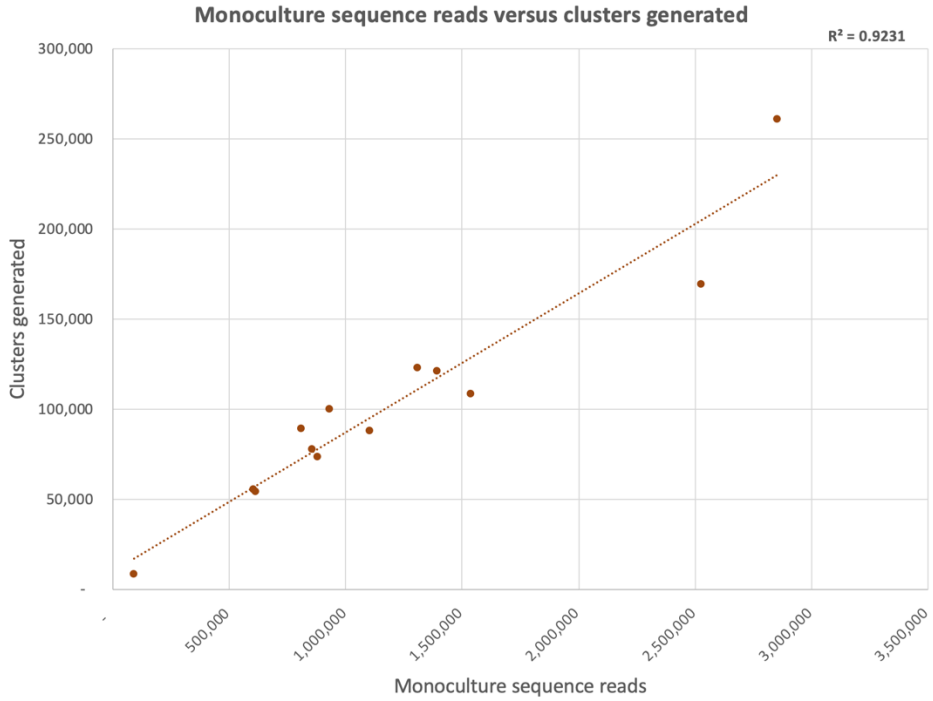


Figure S4 Correlation (R2=0.92) between number of input reads and the number of clusters generated (before filtering) for all 13 monocultures of the *Jena field study*. For most species more sequencing effort will result in more clusters.

Additional read mapping results

For the *Jena field study* mapping failure of 27.8% and 17.99% of the merged- and joined sequence reads, respectively, was classified as 'alignment too short'. Mapping failure of 38.19% and 38.44% of the merged- and joined sequence reads, respectively, was classified as 'other' which happens if STAR was not able to find good anchor seeds or if sequence reads mapped to repeat regions. STAR cannot find good anchor seeds if a sequence read is not represented in the meta-reference, for example when a cluster was removed during the BLASTN filtering. This suggests that although not a lot of clusters were removed by BLASTN filtering they represented a high number of reads.

MOLECULAR ECOLOGY RESOURCES

Non calibrated analysis

Pool 1	FPS (%)	
Species	**Non-calibrated**	**Calibrated**
F. rubra	0.15	0.18
C. jacea	0.08	0.12
H. pubescens	0.15	0.18
K. arvensis	0.20	0.15
L. vulgare	0.15	1.40
P. lanceolata	0.53	0.66
P. pratense	1.31	0.45
P. pratensis	1.10	0.24
	0.46	**0.42**

Table S8 msGBS mock mixture sample FPS(%) of pool 1; the average percentage of root biomass estimated in non-calibrated and calibrated mode when species were not supposed to be present (actual monoculture biomass was zero).

MOLECULAR ECOLOGY RESOURCES

Calculation of the calibration key

A

	Pool 1	Absolute read counts								
		P. lanceolata	*K. arvensis*	*C. jacea*	*F. rubra*	*H. pubescens*	*P. pratensis*	*L. vulgare*	*P. pratense*	Total read count
Calibration samples Pool 1	1	726	2,299	1,282	1,459	1,398	6,225	1,777	1,665	16,831
	2	984	6,128	3,680	4,893	5,251	21,555	3,178	5,886	51,555
	3	322	2,287	1,627	1,954	2,329	7,664	1,144	6,226	23,553
	4	564	5,171	2,484	4,509	4,687	17,712	2,451	6,622	44,200
	5	220	2,584	1,586	1,651	1,760	8,457	1,855	3,305	21,418
	6	316	3,893	2,352	2,057	2,786	10,471	2,113	3,874	27,862
	7	396	2,707	1,183	2,095	2,445	10,130	1,121	4,397	24,474
	8	548	4,394	2,283	2,703	2,147	9,852	2,957	3,467	28,351
	9	1,227	9,568	5,220	6,124	3,401	25,368	4,006	8,856	63,770
	10	786	5,917	5,470	4,713	3,494	15,547	2,261	5,559	43,747
Average		609	4,495	2,717	3,216	2,970	13,298	2,286	4,986	

B

	Pool 1	Relative read counts							
		P. lanceolata	*K. arvensis*	*C. jacea*	*F. rubra*	*H. pubescens*	*P. pratensis*	*L. vulgare*	*P. pratense*
Calibration samples Pool 1	1	0.0431	0.1366	0.0762	0.0867	0.0831	0.3699	0.1056	0.0989
	2	0.0191	0.1189	0.0714	0.0949	0.1019	0.4181	0.0616	0.1142
	3	0.0137	0.0971	0.0691	0.0830	0.0989	0.3254	0.0486	0.2643
	4	0.0128	0.1170	0.0562	0.1020	0.1060	0.4007	0.0555	0.1498
	5	0.0103	0.1206	0.0740	0.0771	0.0822	0.3949	0.0866	0.1543
	6	0.0113	0.1397	0.0844	0.0738	0.1000	0.3758	0.0758	0.1390
	7	0.0162	0.1106	0.0483	0.0856	0.0999	0.4139	0.0458	0.1797
	8	0.0193	0.1550	0.0805	0.0953	0.0757	0.3475	0.1043	0.1223
	9	0.0192	0.1500	0.0819	0.0960	0.0533	0.3978	0.0628	0.1389
	10	0.0180	0.1353	0.1250	0.1077	0.0799	0.3554	0.0517	0.1271
Average = calibration key		0.0155	0.1271	0.0768	0.0906	0.0886	0.3811	0.0659	0.1544
Standard deviation		0.0034	0.0180	0.0204	0.0108	0.0162	0.0302	0.0184	0.0430
CV (%)		21.8	14.2	26.6	11.9	18.3	7.9	27.9	27.8

Table S9 The absolute and relative to total read counts of the calibration samples of pool 1. The absolute read counts (A) vary a lot across species. The relative to total read counts (B) of the calibration samples which, if averaged result in the 'calibration key'. The relative to total read counts vary between species but follow the same pattern across calibration samples. The coefficient of variation (CV), after excluding outlier sample 1 (of which, 1 or more per species values, deviated more than >2.5STD), varies between 7.9 – 27.9%.

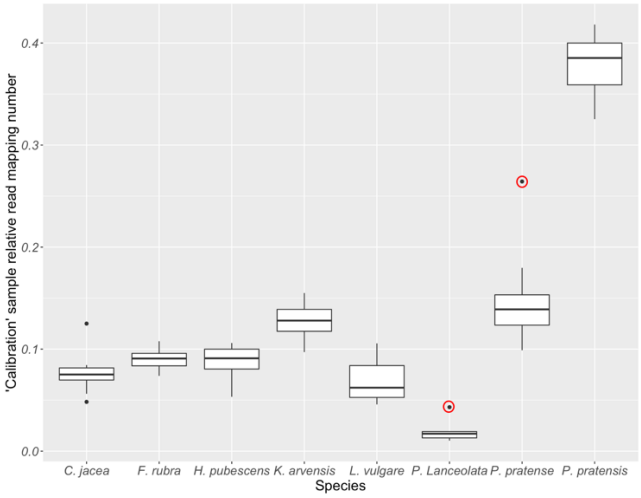


Figure S5 Standard boxplot of the relative mapped read counts of the calibration samples of pool 1 showing the outlier values (>2.5STD) of calibration sample 1 within a red circle.

MOLECULAR ECOLOGY RESOURCES

A

Pool 2		Absolute read counts								
		A. odoratum	*D. glomerata*	*G. pratense*	*H. lanatus*	*L. vulgare*	*P. pratense*	*P. lanceolata*	*R. acris*	Total reads
Calibration samples	11	65	787	76	258	589	261	57	420	2,513
	12	277	2,780	297	169	386	632	204	1,005	5,750
	13	1,041	8,723	853	615	1,893	1,701	469	2,616	17,911
	14	562	5,961	544	328	1,588	955	331	1,929	12,198
	15	8,015	9,597	1,039	680	1,411	2,686	563	4,747	28,738
	16	460	4,136	308	1,041	646	1,327	319	2,021	10,258
	17	624	4,566	618	396	1,113	1,622	404	1,791	11,134
	18	12	1,212	596	26	35	381	321	1,046	3,629
	19	113	1,238	132	95	287	373	79	304	2,621
	20	760	6,969	827	514	1,411	1,246	404	3,171	15,302
Average		1,193	4,597	529	412	936	1,118	315	1,905	

B

Pool 2		Relative read counts							
		A. odoratum	*D. glomerata*	*G. pratense*	*H. lanatus*	*L. vulgare*	*P. pratense*	*P. lanceolata*	*R. acris*
Calibration samples	11	0.0259	0.3132	0.0302	0.1027	0.2344	0.1039	0.0227	0.1671
	12	0.0482	0.4835	0.0517	0.0294	0.0671	0.1099	0.0355	0.1748
	13	0.0581	0.4870	0.0476	0.0343	0.1057	0.0950	0.0262	0.1461
	14	0.0461	0.4887	0.0446	0.0269	0.1302	0.0783	0.0271	0.1581
	15	0.2789	0.3339	0.0362	0.0237	0.0491	0.0935	0.0196	0.1652
	16	0.0448	0.4032	0.0300	0.1015	0.0630	0.1294	0.0311	0.1970
	17	0.0560	0.4101	0.0555	0.0356	0.1000	0.1457	0.0363	0.1609
	18	0.0033	0.3340	0.1642	0.0072	0.0096	0.1050	0.0885	0.2882
	19	0.0431	0.4723	0.0504	0.0362	0.1095	0.1423	0.0301	0.1160
	20	0.0497	0.4554	0.0540	0.0336	0.0922	0.0814	0.0264	0.2072
Average = calibration key		**0.0465**	**0.4392**	**0.0455**	**0.0500**	**0.1128**	**0.1107**	**0.0294**	**0.1659**
Standard deviation		0.0092	0.0571	0.0094	0.0302	0.0504	0.0244	0.0044	0.0268
CV (%)		19.8	13.0	20.7	60.4	44.7	22.1	15.1	16.1

Table S10 The absolute and relative to total read counts of the calibration samples of pool 2. The absolute read counts (A) vary a lot across species. The relative to total read counts (B) of the calibration samples which, if averaged result in the 'calibration key'. The relative to total read counts vary between species but follow the same pattern across calibration samples. The coefficient of variation (CV), after excluding outlier sample 15 and 18 (of which, 1 or more per species values, deviated more than >2.5STD), varies between 13.0 – 60.4%.

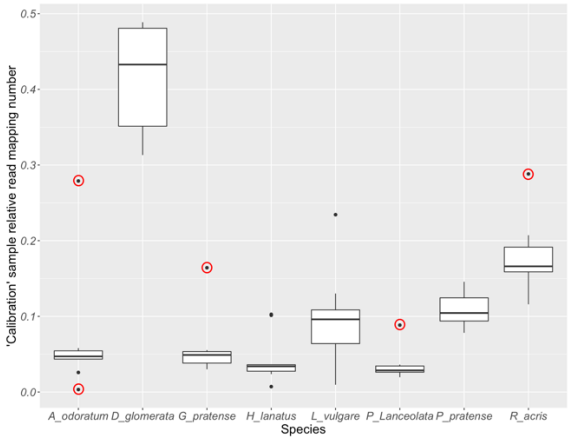


Figure S6 Standard boxplot of the relative mapped sequence read counts of the calibration samples of pool 2 showing the outlier values (>2.5STD) of calibration samples 15 and 18 within a red circle.

Final analysis of the Jena field study

Pool 1	Correlation (R2)			Slope		
Species	Biomass versus qPCR	Biomass versus msGBS	qPCR versus msGBS	Biomass versus qPCR	Biomass versus msGBS	qPCR versus msGBS
F. rubra	0.97	0.98	0.99	1.09	1.03	0.94
C. jacea	0.97	0.95	0.97	1.03	0.92	0.90
H. pubescens	0.98	0.95	0.97	0.93	0.94	1.01
P. pratensis	0.98	0.98	0.99	0.91	1.00	1.09
K. arvensis	0.98	0.97	0.99	1.03	0.98	0.96
L. vulgare	0.99	0.98	0.97	1.05	1.04	0.99
P. lanceolata	0.97	0.97	0.97	1.03	0.92	0.99
P. pratense	0.94	0.97	0.95	0.81	0.89	1.05

Table S11 Correlations (R2) of root biomass versus qPCR, root biomass versus msGBS and qPCR vs msGBS, respectively, of mock mixture root samples of pool 1. For slopes depicted in green, the slope=1 line was within the 0.95 confidence interval boundaries.

Pool 2	Correlation (R2)			Slope		
Species	Biomass versus qPCR	Biomass versus msGBS	qPCR versus msGBS	Biomass versus qPCR	Biomass versus msGBS	qPCR versus msGBS
L. vulgare	0.98	0.97	0.98	0.94	0.85	1.08
P.lanceolata	0.94	0.91	0.92	0.87	0.71	1.16
P. pratense	0.98	0.97	0.97	1.11	1.04	1.05
A. odoratum	0.99	0.95	0.92	0.99	0.99	0.93
D. glomeratum	0.98	0.97	0.95	0.90	0.94	0.91
G. pratense	0.94	0.95	0.98	0.84	0.95	0.88
H. lanatus	0.97	0.85	0.85	0.92	0.64	1.22
R. acris	0.96	0.96	0.95	1.14	1.03	1.09

Table S12 Correlations (R2) of root biomass versus qPCR, root biomass versus msGBS and qPCR vs msGBS, respectively, of mock mixture root samples of pool 2. For slopes depicted in green, the slope=1 line was within the 0.95 confidence interval boundaries.

MOLECULAR ECOLOGY RESOURCES

Pool 1	FPS (%)	
Species	qPCR	msGBS
F. rubra	0.24	0.18
C. jacea	0.05	0.12
A. pubescens	0.11	0.18
P. pratensis	0.28	0.24
K. arvensis	0.03	0.15
L. vulgare	0.24	1.40
P. lanceolata	2.18	0.66
P. pratense	0.66	0.45
	0.47	**0.42**

Table S13 Mock mixture sample FPS(%) for pool 1; the average percentage of root biomass estimated when the species was supposed not to be present (actual monoculture biomass was zero), by qPCR and msGBS.

Pool 2	FPS (%)	
Species	qPCR	msGBS
L. vulgare	0.04	0.61
P. lanceolata	1.57	4.56
P. pratense	0.35	1.02
A. odoratum	0.09	0.30
D. glomeratum	0.16	0.02
G. pratense	0.03	0.02
H. lanatus	0.23	3.51
R. acris	0.67	0.72
	0.39	**1.35**

Table S14 Mock mixture sample FPS(%) for pool 2; the average percentage of root biomass estimated when the species was supposed not to be present (actual monoculture biomass was zero), by qPCR and msGBS.

MOLECULAR ECOLOGY
RESOURCES

A

Pool 1	Correlation (R2) qPCR vs msGBS	
Species	**filtering**	**no filtering**
F. rubra	0.99	0.97
C. jacea	0.97	0.93
H. pubescens	0.97	0.96
P. pratensis	0.99	0.93
K. arvensis	0.99	0.96
L. vulgare	0.97	0.96
P. lanceolata	0.97	0.94
P. pratense	0.95	0.98
Average	**0.98**	**0.95**

B

Pool 1	msGBS - FPS (%)	
Species	**filtering**	**no filtering**
F. rubra	0.18	1.85
C. jacea	0.12	2.92
A. pubescens	0.18	0.64
P. pratensis	0.24	0.78
K. arvensis	0.15	1.47
L. vulgare	1.40	0.54
P. lanceolata	0.66	5.57
P. pratense	0.45	2.66
Average	**0.42**	**2.05**

Table S15 The combined effect of the BLASTN- and monoculture mapping based cluster filtering on (A) the correlation of qPCR versus msGBS and (B) the msGBS FPS(%) within species pool 1.

MOLECULAR ECOLOGY
RESOURCES

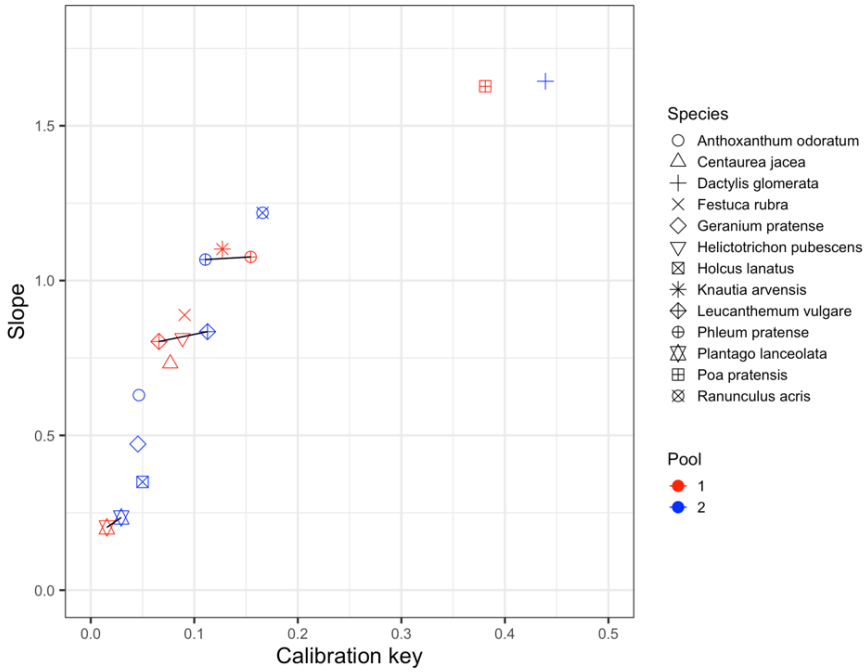


Figure S7 The influence of the species assembly on the calibration key values (Table S5B and S6B) and the non-calibrated regression model slopes (Fig 4) of pool 1 (red) and pool 2 (blue). Symbols of species present in both pools (*Plantago*, *Leucanthemum* and *Phleum*) are connected by black lines. The non-calibrated regression model slopes (of the biomass-based species proportions and non-calibrated msGBS estimated within-species abundance) of the three species present in both pools is hardly altered by the species assembly. For *Leucanthemum vulgare* and *Phleum pratense* the calibration key is shifted, possibly due to the different species assembly.

MOLECULAR ECOLOGY
RESOURCES

Species	Pool 1		Pool 2	
	Slope	Calibration_key	Slope	Calibration_key
Centaurea jacea	0.73	0.08		
Festuca rubra	0.89	0.09		
Helictotrichon pubescens	0.81	0.09		
Knautia arvensis	1.10	0.13		
Leucanthemum vulgare	0.80	0.07	0.83	0.11
Phleum pratense	1.08	0.15	1.07	0.11
Plantago lanceolata	0.20	0.02	0.23	0.03
Poa pratensis	1.63	0.38		
Anthoxanthum odoratum			0.63	0.05
Dactylis glomerata			**1.64**	**0.44**
Geranium pratense			0.47	0.05
Holcus lanatus			0.35	0.05
Ranunculus acris			1.22	0.17

Table S16 The influence of the species assembly on the calibration key values (Table S5B and S6B) and the non-calibrated regression model slopes (Fig 4) of pool 1 and pool 2. The slope and calibration keys of species present in both pools (*Plantago*, *Leucanthemum* and *Phleum*) were compared. The per species shift of the slope ranged from 0.01 to 0.03. The per species shift of the calibration key ranged from 0.01 to 0.4. The slope is less influenced by species assembly compared to the calibration key. However, in relation to other species, the calibration key shift is limited.

MOLECULAR ECOLOGY
RESOURCES

Final analysis of the Dutch Field Study

To illustrate the data behind the rFPS analysis within congener species groups we included Figure S8 and S9. All monocultures of the *Dutch field study* were represented by a robust set of clusters in the meta-reference. msGBS signals of species not expected to be present within a sample (based on extensive field surveys) are assumed to derive from congener species. This assumption is expected to cause a comparable rFPS between samples. However this was not always observed (for examples as discussed for *Cirsium*). Also in the Ranunculus example below the rFPS is not always present. It can be discussed if the rFPS might also not be caused by hybrid species. However for the closely related congener pair *Ranunculus bulbosus* and *Ranunculus repens* (Figure S8) no hybrids are reported in literature (Coles, 1973). For the closely related congener pair *Rumex acetose* and *Rumex thysiflorus* (Figure S9) hybrids are reported in literature (Świetlińska, 2015).

MOLECULAR ECOLOGY
RESOURCES

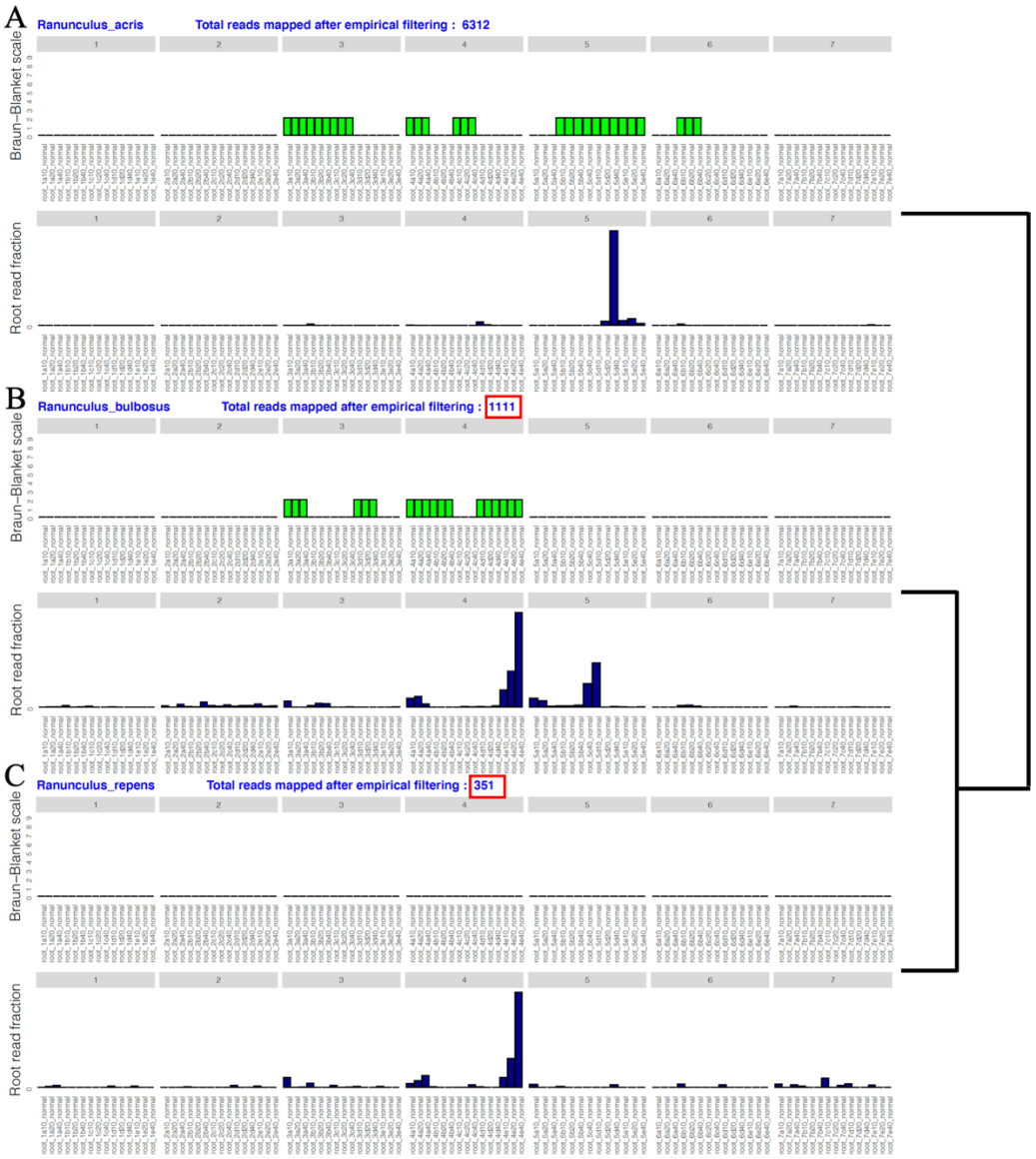


Figure S8 Congener analysis. Example of the *Ranunculus spp.* Field survey of plot 1-7 (ABC-upper; in numeric 0-9 Braun-Blanquet scale) and msGBS (ABC-lower) within species abundance of 7 field locations, 5 plots per location and 3 soil core depths per plot. The phylogenetic relatedness is plotted to the right of the graphs. The *Ranunculus acris* msGBS signal leaves no FPS imprint in the *Ranunculus bulbosus* and *Ranunculus repens* msGBS signal, this correspond to their phylogenetic relatedness (Schuster et al., 2015). It looks as if *Ranunculus bulbosus* derived reads leaves an imprint (44% on average) in the *Ranunculus repens* msGBS analysis, this correspond to their phylogenetic relatedness. However, the *Ranunculus bulbosus* msGBS signal of location 5 subplot c and d did not leave an imprint in *Ranunculus repens* and did also not correspond to the field survey data. It is of course a possibility that a seedling, a seed or an underground root organ is missed during the field survey.

MOLECULAR ECOLOGY
RESOURCES

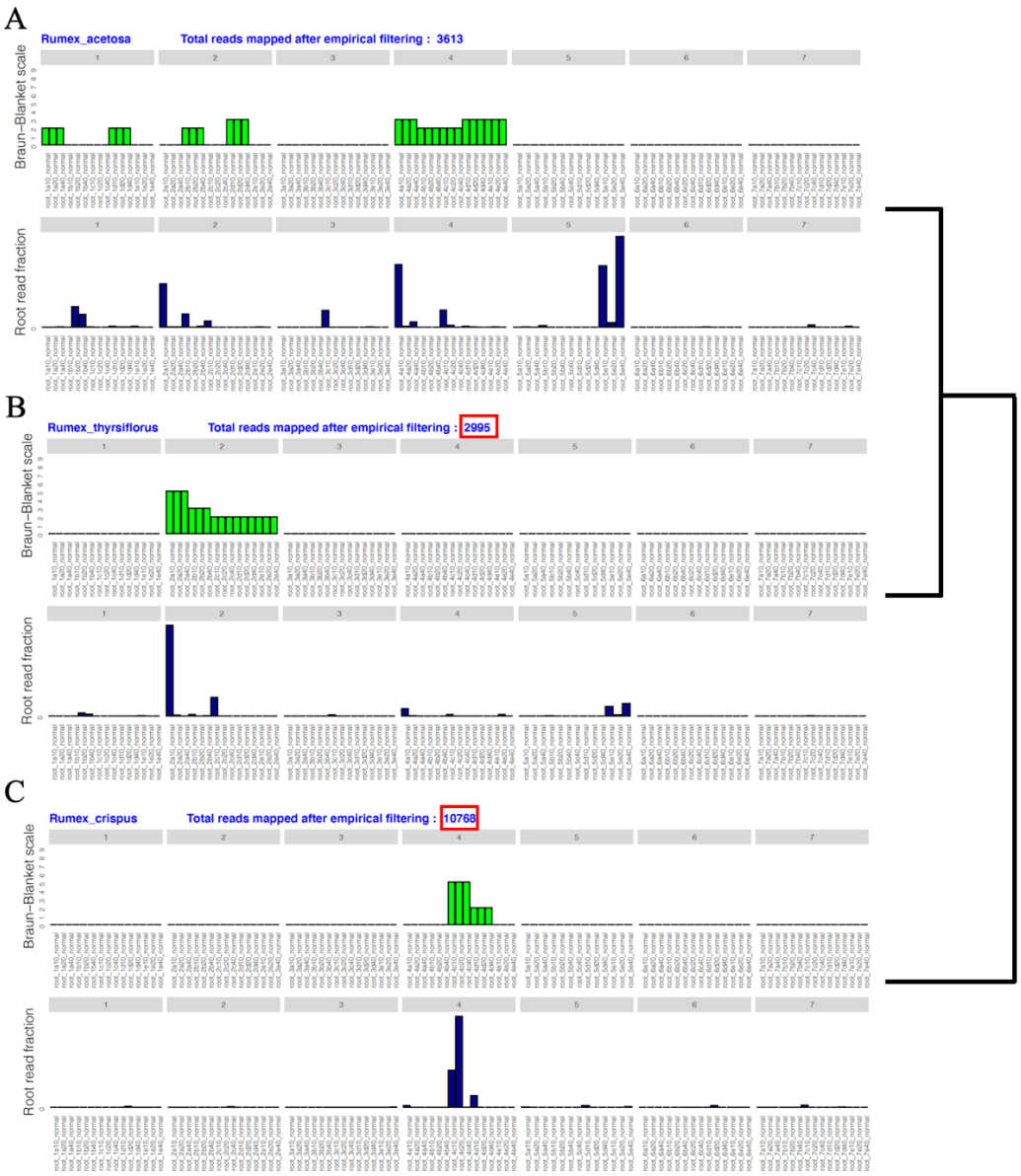


Figure S9 Congener analysis. Example of the *Rumex spp.* Field survey of plot 1-7 (ABC-upper; in numeric 0-9 Braun-Blanquet scale) and msGBS (ABC-lower) within species abundance of 7 field locations, 5 plots per location and 3 soil core depths per plot. The phylogenetic relatedness is plotted to the right of the graphs. The *Rumex acetosa* and *Rumex thyriflorus* msGBS signals leave no FPS imprint in the *Rumex crispus* msGBS signal, this correspond to their phylogenetic relatedness (Baltisberger & Hörandl, 2016). It looks as if *Rumex acetosa* derived reads leaves an imprint (31% on average) in the *Rumex thyriflorus* msGBS analysis, this correspond to their phylogenetic relatedness. However, for example, the *Rumex acetosa* msGBS signal of location 5 subplot e did not correspond to the field survey data. It is of course a possibility that a seedling, a seed or an underground root organ is missed during the field survey.

MOLECULAR ECOLOGY
RESOURCES

Data availability

NCBI Sequence read Archive (SRA)

Raw sequence data can be found at BioProject PRJNA604964:
https://www.ncbi.nlm.nih.gov/bioproject?term=PRJNA604964&cmd=DetailsSearch

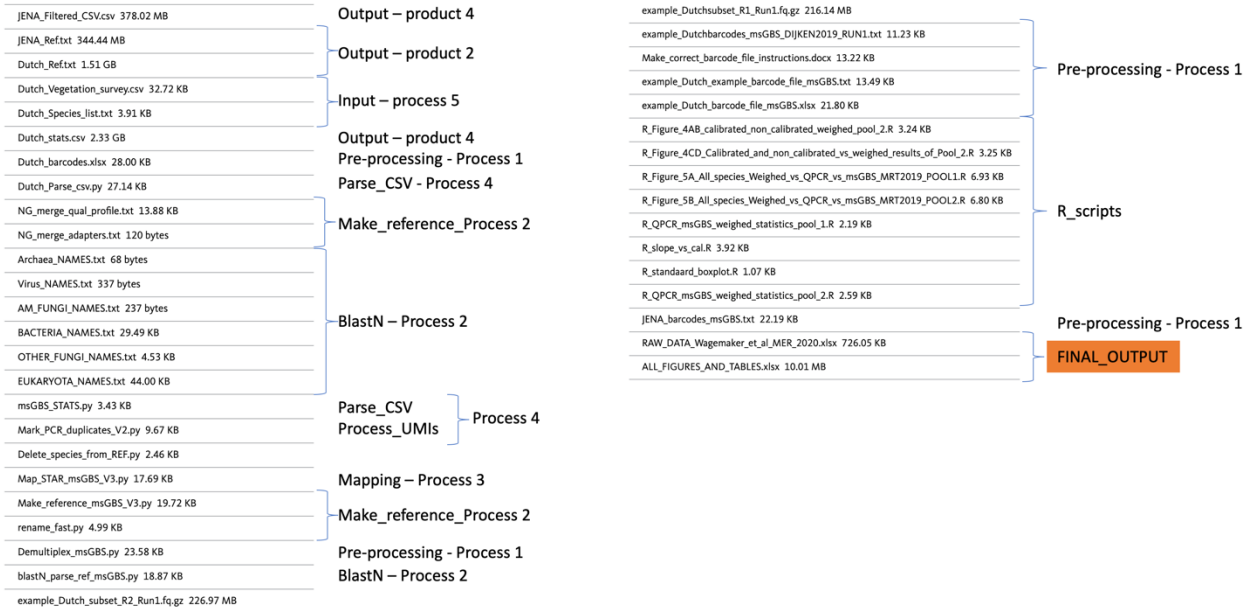
Github

All scripts can be found at:
https://github.com/NielsWagemaker/scripts_msGBS/tree/msGBS-1.0

Dryad

All metadata can be found at:
https://doi:10.5061/dryad.m63xsj3xz

Overview of all available data:



References

Anaconda. (2016). *Anaconda*. Retrieved from https://anaconda.com

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

Burrows, M., & Wheeler, D. J. (1994). A Block-sorting Lossless Data Compression Algortihm. *SRC Research Report*.

Coles, S. M. (1973). Ranunculus bulbosus L . in Europe. *Watsonia*, *9*, 207–228.

Gaspar, J. M. (2018). NGmerge: Merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics*, *19*(1), 1–9. https://doi.org/10.1186/s12859-018-2579-2

Köster, J., & Rahmann, S. (2012). Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

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

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

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*(1), 1–7. https://doi.org/10.1186/s13104-016-1900-2

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

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

Świetlińska, Z. (2015). Cytogenetic relationships among Rumex acetosa, Rumex arifolius and Rumex thyrsiflorus. *Acta Societatis Botanicorum Poloniae*, *32*(2), 215–279. https://doi.org/10.5586/asbp.1963.012

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2017). Genome analysis Sambamba : fast processing of NGS alignment formats. *Bioinformatics*, *31*(November), 2032–2034. https://doi.org/10.5281/zenodo.13200.Contact

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

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

Van Gurp, T. P., Wagemaker, N. C. A. M., Wouters, B., Vergeer, P., Ouborg, J. N. J., & Verhoeven, K. J. F. (2016). EpiGBS: Reference-free reduced representation bisulfite sequencing. *Nature Methods*, *13*(4), 322–324. https://doi.org/10.1038/nmeth.3763

Zaleski, C., Jha, S., Gingeras, T. R., Batut, P., Davis, C. A., Chaisson, M., … Drenkow, J. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635