

Running title: Improving terrestrial eDNA methods

Aligning terrestrial eDNA sampling and analytical choices to effectively capture both rare species and compositional variation in grassland plant communities

Jan Plue ^{1,*}, † & Mats Töpel^{1,*}, †

¹ Swedish Biodiversity Centre, SLU Swedish Species Information Centre, Swedish University for Agricultural Sciences, Uppsala, Sweden

† Both authors contributed equally to this work.

Correspondence: Jan Plue (ORCID 0000-0002-6999-669X); E-mail: jan.plue@slu.se, and Mats Töpel (ORCID 0000-0001-7989-696X); E-mail: mats.topel@slu.se

Funding information

MT and JP received funding from the Swedish Institute for Water and Air Pollution Research (SIVL) for this work. JP also held grants by FORMAS (nr 2018-00961, 2022-01803), and MT was part of the Mistra Finbio research programme (Dnr 2020/10).

Key words (max. 8) Environmental DNA; *Gentianella campestris*; Metabarcoding; Methodology; PacBio sequencing; Plant community composition; Representative; Semi-natural grasslands

Acknowledgements

Mikaela Boltenstern is acknowledged for her field- and lab-work assistance during the project. The authors would also like to acknowledge support of the National Genomics Infrastructure (NGI) / Uppsala Genome Center and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure. Work performed at NGI / Uppsala Genome Center has been funded by RFI / VR and Science for Life Laboratory, Sweden. The molecular biology analysis of the soil samples were performed by the laboratory at IVL Swedish Environmental Research Institute, Stockholm. The authors have no conflict-of-interest to declare.

Data availability statement

Upon acceptance, all raw data will be archived in an online Dryad digital repository (Plue & Töpel, 2026).

Abstract (300 words max.)

Vascular plants are a major component of terrestrial diversity, yet they are overrepresented among the world's threatened species. To effectively manage this biodiversity crisis, data with high spatiotemporal resolution are crucial, yet often lacking for plants. Environmental DNA analysis (eDNA) is capable of rapid detection of biodiversity by metabarcoding the collection of DNA molecules retrieved from environmental samples such as soil cores. The technology may soon support the generation of broad-scale longitudinal plant community data, yet much methodological work on sampling strategies and analytical choices remains if soil-based eDNA is to become a reliable tool for monitoring terrestrial plant communities. Therefore, this dual purpose study in seven Swedish semi-natural grasslands investigated if and when eDNA-generated community data can be used as a stand-alone information source 1) to inform on the presence of a rare, small-statured grassland specialist (*Gentianella campestris*) and 2) to simultaneously infer community compositional change. We demonstrate eDNA to be an effective means of finding a rare species in a highly taxonomically diverse habitat, uncovering *G. campestris* DNA in 31% of the core samples. Evidence suggests the eDNA signal reflects recent spatio-temporal population dynamics at fine spatial scales. Although the entire plant community was not uncovered, molecular community data proved a representative subset, effectively capturing changes in community diversity and composition at plot sizes commonly used for plant surveys. Choices surrounding typical RRA-filtering had significant bearing on eDNA's discriminating power: filtering may overly conservatively remove true observations of a rare species, while filtering highly localized plot noise led to more robust patterns emerging in species richness and plant composition turn-over. Given careful alignment of study goals and sampling strategies, soil-based eDNA may already provide a stand-alone tool for generating reliable, scalable and observer-independent longitudinal data for unveiling and monitoring changes in plant diversity in terrestrial habitats.

Introduction

Vascular plants are a major component of terrestrial biomass and diversity (Bar-On et al. 2018; Kier et al. 2005), being crucial for biodiversity at higher and lower trophic levels (Wan et al. 2020), ecosystem processes (e.g. carbon sequestration, Chen et al. 2020), ecosystem stability (Haddad et al. 2011) and ecosystem service provisioning (Letourneau et al. 2011). At the same time, plants are clearly overrepresented among species at risk of extinction (ca. 40% of assessed species across all taxa, Nic Lughadha et al. 2020). A range of global change drivers such as habitat loss, climate change and eutrophication are rapidly transforming and/or eroding plant community diversity at local, regional and global scales (Vellend et al. 2013, Nic Lughadha et al. 2020, Kimmel et al. 2022, Montràs-Janer et al. 2024), underpinning repeated, urgent calls to hasten and improve upon habitat restoration and conservation efforts (Banks-Leite et al. 2020, Watson et al. 2016). However, to identify and effectively manage biodiversity during this conservation crisis, access to data with high temporal and spatial resolution is crucial (Rodrigues et al. 2006), yet lacking in vascular plants (Bachman et al. 2019).

Environmental DNA (eDNA) is a molecular technology capable of detecting biodiversity via the identification of species specific DNA molecules contained within any type of environmental or biological sample. The full potential of ever accelerating eDNA technology may soon be unleashed to rapidly generate broad-scale longitudinal data in terrestrial plant biodiversity monitoring (Yoccoz et al. 2012), particularly as terrestrial plant eDNA applications have shown great promise in a range of habitats (tundra, Edwards et al. 2018; desert, Carrasco-Puga et al. 2021; grassland, Hartvig et al. 2021; boreal forest, Ariza et al. 2023; mediterranean and tropical communities, Duley et al. 2023). Still, while being a key tool of future plant community research and monitoring, much field-, lab- and bioinformatic methodological development remains if terrestrial eDNA is to become a reliable stand-alone tool to perhaps one day replace traditional plant community surveys otherwise hampered by high human-resource and monetary costs, temporal sampling constraints (Yoccoz et al. 2012) and observer bias (Verheyen et al. 2018).

Any observational study or monitoring scheme is only as good as its sampling design and sample representation, mainly guided by the study's objectives. In other words, how well an individual sample reflects the local extant presence of focal species, taxa and/or communities, will determine the chances of capturing subtle spatio-temporal species distributional or community compositional trends across multiple locations. Sampling invisible, below-ground phenomena, *in casu* DNA distributed in the soil, thus presents a formidable challenge to acquire an accurate, representative local sample (cf. Baker et al. 2009; Muukkonen et al. 2009; Plue & Hermy 2012). This must be properly addressed during eDNA soil sampling for plot-based plant community assays, if firm ecological conclusions are to be drawn from the eDNA data (Zinger et al. 2019). The challenges for representative sampling are indeed numerous. First, established plants' intra-and interspecific spatial distributions are spatially clustered in response to environmental heterogeneity (Ulrich et al. 2021), triggering similar belowground patterns (roots, Pärtel & Wilson 2002; seeds, Plue & Hermy 2012; pollen, Pardoe 1996), shaping the soil distribution of local intra-

and extracellular DNA (see Levy-Booth et al. 2007). High soil DNA heterogeneity in itself thus presents a significant obstacle when extracting a single plot-based sample representative of the extant plant community. Second, plant soil eDNA generally contains and accumulates DNA of foreign origin (both spatial and temporal, Levy-Booth et al. 2007), coming from e.g. far-reaching roots (Hiiesalu et al. 2012), pollen dispersal (de Vries et al. 2003), temporal seed dispersal (Plue & Cousins 2018), the persistence of intercellular (e.g. leaf and root litter) and extracellular DNA (adhered to soil complex) of plants no longer present locally (Levy-Booth et al. 2007, Yoccoz et al. 2012, Carini et al. 2016, Ariza et al. 2023) and/or extracellular DNA transported within the soil matrix (Poté et al. 2007). Each process on its own may reduce the chances of acquiring a reliable local, plot-based signal, potentially distorting true plant community patterns (cf. Carini et al. 2016; but see for high eDNA–community taxonomic overlap, Edwards et al. 2018; Ariza et al. 2022). Third, eDNA sample output, in the form of operational taxonomic units (OTU's) or amplified sequence variants (ASV's) is further susceptible to methodological and technological biases introduced during the sampling, analysis and bioinformatics process. In that context, particularly low sequence read abundances are often seen as problematic. Even if low read abundance can result from 1) a true representation of a rare species, 2) poor amplification of an abundant species (Elbrecht & Leese 2015), 3) low abundances of DNA foreign to the plot (cf. earlier), 4) variation in sample sequencing depth (Deagle et al. 2019) or 5) stochasticity during (sub)sampling, DNA extractions, PCRs or DNA-sequencing (Luo et al. 2023), the perceived risk of field- or lab-contamination producing false positives often leads rare reads to be conservatively removed. However, eDNA plant community studies have mostly focused on comparatively species-poor (Edwards et al. 2018; Ariza et al. 2023) rather than species-rich communities, where the majority of plant species are low abundant (Ulrich et al. 2010), and therefore the amount of eDNA they shed and thus their detection likelihood also is low (Yoccoz et al. 2012; Edwards et al. 2018). Yet, low abundance sequence reads may hold invaluable information when applying eDNA technology to trace rare, threatened species (Hartvig et al. 2021), detect the onset of species invasions (Sepp et al. 2021), evaluating a site's restoration potential or assessing the plant community beyond the spatial and temporal limits of the local plot sample (Carrasco-Puga et al. 2021). Consequently, besides mere sampling design considerations per se, post-hoc data curation to reduce methodological and technological bias needs to make informed choices based on the study objectives as these choices may otherwise profoundly impact ecological interpretation of signals in eDNA data.

Taken together, translating raw DNA sequence reads to species occurrence data can be challenging overall, but seems particularly complex when aiming to primarily retain a local plot signal for plant communities. This study therefore aims to identify under what parameters species and community plot-level data detected by eDNA can be used as a stand alone information source to accurately inform on 1) the detection of rare, threatened grassland specialist species for restoration purposes and 2) plot-based plant community composition to infer compositional changes between plots. For those purposes, we designed a dual-purpose study in seven species-rich semi-natural grassland communities holding the rare

134 grassland specialist species *Gentianella campestris*, to investigate: 1) under what sampling regime terrestrial
 135 eDNA can successfully detect the presence of *G. campestris* DNA, 2) at what plot size a plant community
 136 is accurately represented by a single eDNA soil sample, 3) at what plot size do eDNA plant community
 137 data reliably uncover plant compositional changes between plots and 4) how does post-hoc data curation
 138 removing low read abundances affects eDNA species and compositional signals in 1) - 3).

139

140

Materials and Methods

2.1 Study area

We performed the study in seven managed semi-natural grasslands scattered across northern Stockholm county in central Sweden. Nordic semi-natural grasslands are known for their exceptionally high species diversity at small scales due to a long history of continuous grassland management (0.1 – 1 m²; Wilson et al. 2012). Because of severe global declines (e.g. Auffret et al. 2018), semi-natural grasslands are today of global conservation concern (Watson et al. 2016) triggering ongoing efforts to monitor their biodiversity over time (National Inventory of Landscapes in Sweden, Ståhl et al. 2010; National Meadow and Pasture Inventory, <http://www.sjv.se/tuva>), rendering them particularly suitable for our study's main goals. All grasslands, ranging in sizes between 4011 and 128 423 m² (mean of 45 554 ± 15 516 m²) were managed by either mowing (two sites) or livestock grazing by cattle (three sites) or sheep (two sites). A final selection criteria for the study sites was the recorded and geo-located presence of the rare, red-listed grassland specialist species *Gentianella campestris* L., with selected grasslands representing a gradient in the time since *G. campestris* was last observed and a gradient in population size.

2.2 Data collection

eDNA samples

During October 2021, we collected individual soil cores with a sampling scheme which would help us (1) identify the sensitivity of eDNA at detecting *G. campestris*, and how sampling and population factors affect that sensitivity and (2) enable us to investigate if and at what spatial scales (i.e. those generally applied during traditional plant community inventories) eDNA core samples would effectively capture variation in plant community and diversity. Based on available GPS coordinates extracted from the Swedish Species Observation System (Artportalen; www.artportalen.se), we located where *G. campestris* was last observed within each grassland. We then visited these coordinates and collected a total of nine soil cores from each grassland, i.e. three samples randomly spread out in each of three circular bands around these coordinates (0 – 5 m, 5 – 40 m, 40 – 120 m). The coordinates of each sampling point were documented using a high performance GPS (Emlid Reach RS2 RTK) to an accuracy of 10 cm. The soil auger (3.5 cm diameter) was thoroughly cleaned between each soil core sampling event. First, we used a dish brush with de-ionized water to clean the auger of any visible soil and plant material. The auger was then rinsed profusely twice, first with de-ionized water followed by a 5% chlorine solution, the latter to destroy any remaining DNA. A mineral soil core (i.e. removing the litter layer and plant parts prior to sampling, using single-use medical gloves) was then collected to a depth of 5 cm, and placed directly into a plastic zip-lock bag. Soil core samples were immediately put in a cool box and subsequently in a -21°C freezer at the end of each day of fieldwork, ensuring minimal impact on eDNA sample quality (Frøslev et al. 2023).

Sampling and population information

While collecting soil cores in 2021, we equally recorded the distance from each soil core to the nearest *G. campestris* individual. At the same time, we counted all *G. campestris* individuals in a 2 m radius from the soil

core to estimate 2021 population density. Upon returning for traditional plant community inventories in 2022, we again counted all *G. campestris* individuals within a 2 m radius of each core sample location to estimate the 2022 population density. Finally, we extracted population size data from the Swedish Species Observation System, where designated flora guards (SE Floraväktarna; [Website](#)) regularly report population sizes of red-listed species from locations to which they were assigned. As such, we have exact data on *G. campestris* population size during the last eight years, as well as information on the year when the species was last visually observed. As the species is biennial with significant inter-annual demographical variation (Lennartsson & Oostermeijer 2001), we used the eight year population size average as a population size proxy.

Plant community inventories

In June 2022, we used the same high accuracy GPS (Emlid Reach RS2 RTK) to return to each of the 63 sampling locations. For 62 samples, we managed to re-locate the exact hole which was left behind after autumn core sampling. We inventoried plant communities centered around the core sample location along a nested plot design. Plot size increased from 10 × 10, to 25 × 25, 50 × 50 and 100 × 100 cm² centered around the soil core location. Per plot size, we recorded the presence and cover [%] of all vascular plant species, returning a total of 252 plant community inventories.

2.3 DNA extraction and PCR

Soil samples were transferred to 50 ml Falcon tubes and then freeze dried for approximately 24 hours. Glass beads were then added and the samples were homogenized using a FastPrep-24 5G (MP Biomedicals, LLC). DNA was extracted using the DNeasy power soil pro kit (Qiagen) according to the manufacturer's instructions. DNA-concentration was then quantified using Qubit™ 1X dsDNA High Sensitivity (HS) assay kit in the Qubit Flex fluorometer (ThermoFisher).

DNA concentration of each sample was diluted to 10 ng/μl after which the PCR reactions were performed in 25μl volumes containing 2μl template DNA, 12.5 μl 2X Phusion™ Plus Green PCR Master Mix, 1.25 μl of 10 μM forward and reverse primer solution mixes (see Table S2 in Supplementary Materials) and 8 μl DNase free water. The PCR-program started with a 2 min. denaturation at 98°C followed by 5 cycles of 10 min. at 98 °C, 35 sec. at 53 °C and 30 sec. at 72 °C, then 35 cycles of 10 sec, at 98 °C, 30 sec. at 63.6 °C and 30 sec. at 72 °C and finalised with a 5 min. extension period at 72 °C. Each sample was amplified in triplicate and then pooled before being purified using the AMPure magnetic beads assay (Beckman Coulter, US).

This second PCR was performed in a final volume of 25 μL containing 2 μL of purified PCR-product, 5 μL of 5X Q5 reaction buffer, 0.4 mM M13 forward and M13 Reverse index primers, 0.2 mM dNTPs, and 0.025 U/μL Q5 High-Fidelity DNA Polymerase (New England Biolabs), following to the PCR-protocol provided by the sequencing facility SciLifeLab, Uppsala. Amplicons were then purified with SMRTbell shape cleanup beads to remove too short fragments.

2.4 Bioinformatic analysis and quality control

The DNA-library was sequenced on one SMRT® Cell 8M v3 of the PacBio Sequel II platform at the Uppsala Genome Center of the National Genomic Infrastructure, SciLifeLab. The resulting sequences were processed using the nf-core/ampliseq pipeline v2.7.0 (Ewels et al. 2020), that included quality analysis using fastQC (Andrews 2010), trimming and adaptor removal using cutadapt (Martin 2011), sequencing error correction and taxonomic classification of the amplified sequence variants (ASVs) using dada2 (see Table S3 in Supplementary Materials for parameter settings; Callahan et al. 2016). For the taxonomic classification, as custom reference sequence database was used containing 4946 sequences downloaded from the National Center for Biotechnology Infrastructure (<https://www.ncbi.nlm.nih.gov/>) from 2251 species known to occur in the region of Uppland (Sweden) as identified according to Upplands Flora (Jonsell 2010).

The number of raw reads for each sample ranged between 105827-30450 sequences (median 54527) with the exception of four samples with low yields (1100, 121, 79 and 67 reads respectively). These low yield samples were not excluded from the analysis as they did not behave as outliers by any other measure amongst the derived variables, and their omission had no effect on the results. Between 41-83% of the reads remained after quality filtering and adaptor trimming and were subject to taxonomic classification by dada2 (Callahan et al. 2016). In total, 731 ASVs were detected in the 62 soil core samples. Some 622 were taxonomically classified to species level and one only to the *Convallaria* genus. The remaining ASVs were identified as *Streptophyta* (78) or as *Eukaryota* (29). In addition, 69 ASVs had an exact match (via *species_exact* by dada2) to a sequence in the reference database.

Then, we manually curated the raw ASV data for quality control and data cleaning purposes. ASV nucleotide sequences were subject to similarity searches using BLASTn (via BLAST+ v2.14.0; Camacho et al. 2009), querying sequences against the NCBI nucleotide database using default parameters. Following this, a series of changes were made to the automatic bioinformatic classification. First, all ASVs previously only identified as *Streptophyta* could be assigned to the species (71) or genus level (7). Most ASVs were of species already identified via the bioinformatic pipeline, but 13 ASVs were assigned to nine species previously unidentified by the bioinformatic pipeline (*Galium mollugo*, *Geum rivale*, *Myosotis stricta*, *Ranunculus acris*, *R. repens*, *Saxifraga granulata*, *Sesleria caerulea*, *Solidago virgaurea* and *Stellaria alsine*). Second, ASVs only classified as *Eukaryota* proved to be bacteria and were omitted. Third, species assignments to reference sequences representative of multiple congeneric species were assigned to a single species based on either BLASTn assignment or expert judgement based on Swedish species distributions or clear differences in habitat preferences. For example, an ASV assigned to reference sequence KF737610.1 representing both *Sagina procumbens* and *S. saginoides* was identified as *S. procumbens* since *S. saginoides* most southern Swedish population was ca. 400 km to the east of our study grasslands. Fourth, BLASTn identified that two species (*Festuca rubra* and *Galium boreale*) could be further subdivided into *Festuca rubra* and *F. ovina* and *Galium boreale* and *G. verum*. Both the ASVs and the final, manually curated 78 species ×

63 plot data matrix are available on Dryad (Anonymous 2026).

2.5 Data analysis

Data preparation

The primary objective of the study was to evaluate if data on the presence-absence of a rare species (*G. campestris*) as well as data on the plant community unveiled via eDNA analysis can be used as a stand alone information source to accurately inform on plot-based species detection and plant community composition in the established vegetation, inferring species distributional and community compositional changes between plots. To investigate if mitigation of methodological and technological bias in an ASV data matrix is dependent upon study objectives, we used both a full and a reduced species $\times$ sample ASV data matrix (See Datasets S2 and S3; will be deposited at Dryad). The reduced ASV data matrix was calculated from the full data matrix. We used sample-based relative read abundances (RRA) to remove all ASVs detected below a threshold of 0.05% of the total number of ASVs for that sample (Deagle et al. 2019), hypothesizing that this would 1) decrease the likelihood of detecting rare species and 2) increase the likelihood that eDNA plant community data would be a more accurate representation of the established plant community and compositional trends among plots. Finally, both the full and reduced ASV data matrices had ASV read abundances transformed to presence-absence.

Detection likelihood

To investigate if the chance of detecting *G. campestris* DNA was dependent upon sampling and inter-annual variation in population characteristics, the presence/absence of *G. campestris* in the 63 eDNA soil samples was analyzed using mixed effect logistic regressions (*glmer* from *nlme* package, Binomial distribution with *logit* link function). We constructed a full model focusing on testing the combined impact of sampling distance and population characteristics:

$$p(\text{DNA } G. \text{ campestris}) \sim \text{Distance} * \text{Population size} * \text{Population density}_{21|22} + (\sim 1 | \text{site})$$

This model contained the main variables of interest as fixed effects and potential interactions, while grassland site was incorporated as a random intercept to accommodate for confirmed inter-site variation (likelihood ratio $\text{Chi}^2 = 5.54$; $p = 0.02$; Zuur et al. 2009). However, given high collinearity between distance and population density_{21|22} (Spearman $\rho_{21|22} = -0.65$ | -0.61 , $p < 0.001$), we could not reliably combine these variables in a single model. Instead, we added each fixed effect separately to a null model containing only the random intercept term, using likelihood ratio tests (function *anova*) to identify the direct importance of distance, population size and population density (for both 2021 and 2022) separately. In a next step, we tested the additive effect of population size on detection likelihood, by adding this predictor variable to the random effect (*site*) models containing either distance or population density (only 2021) as a fixed effect, as these predictor variables did not co-vary strongly with population size (Spearman $\rho = -0.24$, $p = 0.06$; Spearman $\rho = 0.31$, $p = 0.01$, respectively). Two-way interactions in these models were tested but proved non-significant. All predictor variables were scaled so that

comparable standardized effect sizes (SES) would be returned (Grueber et al. 2011), which allows for comparison of effect sizes of each predictor within and across models. Finally, to investigate the effectiveness of eDNA to trace a rare species and the impact of data curation on sampling effort guidelines for its detection, the above analysis was run on two presence/absence variables of *G. campestris*: 1) using all recorded presences, irrespective of the number of detected ASVs and 2) using only those records with an RRA > 0.05%.

Community representation

To test if, and at what plot size, species richness captured via a single eDNA soil sample differs from the species richness from the vegetation, we calculated species richness per sampling effort (eDNA core sample location = 0, 10 cm distance, 25 cm, 50 cm and 100 cm). Species richness was then analyzed using generalized linear mixed effect models following:

$$Richness \sim Sampling\ effort + (\sim 1 | site/plotID),$$

including *sampling effort* (5 factor levels) as fixed effects and *plotID* nested in *grassland site* as a random intercept, with a Poisson error distribution, respectively (using the *glmer* function in the *nlme* R-package). *plotID* was included to account for the longitudinal, i.e. nested, plant community inventories centered around each soil core. Homogeneous subsets were identified by resetting the base factor level for *sampling effort* used for comparison and re-running the model. Furthermore, to investigate if, and at what plot size, a single eDNA soil sample may most accurately represent extant plant community composition, we calculated the Raup-Crick similarity metric (*vegdist* function in the *vegan* R-package; Oksanen et al. 2023) to assess compositional similarity between the plant community identified in an individual soil sample using eDNA vs. the plant communities inventoried at various plot sizes. For comparison to other studies, we also calculated the Bray-Curtis similarity metric which is prone to severe bias when species richness differs between communities under comparison (Koleff et al. 2003; as observed in this study). Bray-Curtis similarity generally overestimates similarity, a problem avoided via random draws from the entire species pool, rendering Raup-Crick similarity less sensitive to species richness bias and therefore ought to be the preferred metric (Vellend et al. 2007). We used a general linear mixed effect model to compare whether Bray-Curtis similarities were consistently larger than Raup-Crick similarities with *metric* (Bray-Curtis vs. Raup-Crick) as fixed effect and *plotID* (to account for longitudinal sampling) nested in *grassland site* as a random intercept, with a Gaussian error distribution (*glmer* function in the *nlme* R-package).

In order to understand which species may drive observed changing patterns in compositional similarity with plot size, we then calculated three species richness variables per plot size: 1) the number of species unique to the traditional survey data, 2) the number of species unique to the eDNA data and 3) the number of species shared by the eDNA- and traditional survey data. These variables were analyzed following the same generalized linear mixed effect model structure as for species richness, with main differences being 1) *sample effort* comprising only 4 factor levels representing different plot sizes and 2) the

use of Gaussian error distribution for the similarity metrics (square root-transformed for normality). Finally, to assess the impact of data curation on potential patterns of representativeness, the above analyses were rerun using variables calculated based on the species $\times$ sample presence/absence matrix in which all RRAs $< 0.05\%$ had been removed.

Capturing and predicting plant community variation

We finally needed to establish whether eDNA-based plant community data can be reliably used as an accurate, stand-alone indicator of the true spatio-environmental variation in species richness and community composition. Therefore, we first ran general linear models per plot size to test if and at what plot size eDNA species richness would predict species richness observed in the vegetation. Models were run separately with eDNA species richness based on either the full or 0.05% RRA filtered species $\times$ sample matrix to assess how data curation by removing ASVs of low abundance affects discovery in species richness prediction. Secondly, to assert if eDNA plant community data can detect spatio-environmental variation in observed plant community composition we applied non-metric multidimensional scaling to each of five full presence/absence species $\times$ sample matrices (one per sampling effort: eDNA and four plot sizes). For reasons outlined above, Raup-Crick similarities were chosen as the distance metric underpinning the NMDS, whereas we constrained the optimal solution to returning only two axes, rotated to present the main axes of compositional variation in each dataset. For each NMDS, we then extracted the axis scores for both axes for each of the 63 samples. To check whether compositional variation was similar in paired vegetation and eDNA communities, general linear models were then built per plot size where eDNA and vegetation composition (NMDS axes) were the response and predictor variable, respectively. The entire compositional and post-hoc analysis outlined before were subsequently rerun based on the 0.05% RRA filtered species $\times$ sample matrix to assess how data curation affects compositional pattern discovery. All data analyses were done in R v4.2.2 and graphs were created with the *ggplot2* package (R Core Team 2023).

Results

Detection likelihood

DNA of *G. campestris* was found in 20 of the 63 soil samples (31.75%). The species was discovered in six out of seven grasslands, including in two samples from the Åckelsta grassland, collected near the last known location where the species had last been observed eleven years ago, despite annual resurvey efforts. Our logistical models identified that both distance to the nearest individual of *G. campestris* and population characteristics affect *G. campestris*' detection likelihood; in order of importance: 2021 population density (SES = 1.48) $>$ nearest distance ($|SES| = 1.27$) $>$ population size (SES = 0.78). Specifically, both declining distance to the nearest *G. campestris* individual (Model $\chi^2 = 14.00$, $p < 0.001$) as well as increasing population size (Model $\chi^2 = 5.21$, $p = 0.02$) had significant direct, positive effects on the species' detection likelihood. Increasing population density in 2021 significantly elevated the species' detection likelihood (Model $\chi^2 = 12.37$, $p < 0.001$), though the latter was unaffected by

population density in 2022 (Model $\chi^2 = 1.29$, $p = 0.26$). When investigating the additive effects of population size on *G. campestris* detection likelihood, population size did significantly improve the distance model ($\chi^2 = 3.96$, $p = 0.04$), translating in significant effect sizes for both distance (SES = -1.16, $p = 0.002$) and population size (SES = 0.73, $p = 0.03$). The additive effect of population size did improve the 2021 population density model ($\chi^2 = 10.41$, $p = 0.001$), with a positive population density effect (SES = 1.20, $p = 0.02$), but no significant effect of population size (SES = 0.60, $p = 0.06$).

Depending on if RRA filtering was applied, *G. campestris* DNA was only detected in seven out of 63 soil samples (11.11%), and only found in three grasslands. Modeling results were, however, near-identical with direct positive significant impacts on detection likelihood by declining distance to the nearest *G. campestris* individual, increasing 2021 population density and increasing population size, and significant additive effects beyond distance of both population density as well as population size (for full model details, see Table S4 in Supplementary Information). Rather than distance to the nearest individual being most important, SES across different models suggested that population characteristics were more important in explaining *G. campestris*' detection likelihood. Finally, RRA filtering caused the likelihood of detecting *G. campestris* in soil cores taken closest to the last known location of the population, at the highest population densities and in large populations to decline with ca. 40%, 40% and 20%, respectively. Moreover, instead of more gradual changes in detection likelihood, RRA filtering led to rapid exponential declines in detection likelihood with declining population size and increasing distances.

Community representation

The 63 eDNA soil samples and vegetation surveys together yielded a total of 144 unique plant species. A total of 69 species were identified using eDNA, of which three were exclusive to the eDNA analysis: *Agrostis vinealis*, *Helianthemum nummularium* and *Stellaria alsine*. Hence, 47.83% of the observed plant species were uncovered using eDNA. RRA filtering removed 0.13% of all ASV reads, which equally represented the removal of 458 (out of 998; 45.89%) species $\times$ sample occurrences. Moreover, 10 observed species were lost from the original eDNA species $\times$ community matrix (29 unique species $\times$ sample occurrences; 2.91 %), including *A. vinealis* and *H. nummularium*. eDNA species richness in a core sample decreased significantly due to RRA-filtering of the original eDNA data matrix (paired t-test: mean of 15.84 species vs. 8.57 species; $t = 7.27$, $p < 0.001$), yet RRA-filtered eDNA species richness was still correlated with eDNA species richness based on all presence/absences ($r_{\text{Pearson}} = 0.39$, $p = 0.001$).

Using the original full raw read sequence data matrix, eDNA samples contained significantly more species than the extant plant community at the $10 \times 10 \text{ cm}^2$ plot size (SES = -0.36; $p < 0.001$), whereas eDNA samples significantly contained less species than the extant community for the remaining plot sizes (SES_{25cm} = 0.17 < SES_{50cm} = 0.45 < SES_{100cm} = 0.70; $p < 0.001$). After RRA filtering, eDNA samples consistently and significantly contained less species than the extant community at each plot size ($p < 0.001$), yet with a steadily increasing difference in species richness with plot size (SES_{10cm} = 0.24 < SES_{25cm} = 1.00 < SES_{50cm} = 1.59 < SES_{100cm} = 2.26). In both instances, these patterns seem mainly driven

by significantly increasing numbers of species unique to the extant communities (red boxplots in Fig. 4), i.e. species not detected in the eDNA sample. At the same time, species numbers detected in both the eDNA sample and the extant community (recorded by the traditional survey) significantly increased with plot size (blue boxplots in Fig. 4). This may be a result of species present at larger plot sizes were successfully uncovered by the eDNA analysis, as these species were actually being detected in the extant community. Consequently, a significant, gradual decrease in species numbers only detected by eDNA was also detected (green boxplots in Fig. 4). Still, RRA filtering caused less shared species to be uncovered, implying true detections of present species were being excluded due to a common data curation technique.

Compositional similarity between an eDNA core sample and its corresponding plot's plant community was significantly affected by the choice of metric, with the Raup-Crick metric significantly reduced by ca. 2/3 compared to the typically-used Bray-Curtis metric, irrespective whether the eDNA data were RRA filtered or not (Fig. 3a). The suspected species richness bias underlying Bray-Curtis overestimation was underlined when RRA filtering was applied on the eDNA matrix. Species removal from the eDNA matrix significantly *increased* Bray-Curtis compositional similarity ($t_{\text{paired}} = -8.68, p < 0.001$), as opposed to significantly reducing Raup-Crick similarity ($t_{\text{paired}} = 9.74, p < 0.001$), which corrects for differences in species richness (Fig. 3a).

The Raup-Crick compositional similarity between an eDNA core and a plot's plant community was highest at the $10 \times 10 \text{ cm}^2$ plot size (mean $41 \pm 34\%$), which was significantly higher than eDNAs compositional similarity with the plant community at all other plot sizes ($\text{SES}_{25\text{cm}|50\text{cm}} = -0.13, \text{SES}_{100\text{cm}} = -0.11; p < 0.001$). Compositional similarity did not differ significantly amongst the remaining plot sizes ($27 \pm 24\% - 31 \pm 27\%; p > 0.05$).

Capturing and predicting plant community variation

eDNA species richness based on the full presence/absence data matrix was unable to predict species richness observed in the extant plant community at any plot size (Fig. 5 a-d). After applying a $<0.05\%$ RRA filter, eDNA species richness could significantly predict species richness of the extant plant community at 50 cm and 100 cm plot sizes ($p = 0.02$ and 0.01 , respectively), but not at the small plot sizes (10 cm and 25 cm; $p = 0.07$ and 0.07 , respectively; Fig. 5 e-h). eDNA samples could thus reliably discern between vegetation plots with low vs. high species richness in plot sizes commonly applied in grassland plant inventories when low read abundances were removed from the observed eDNA matrix.

All NMDS ordinations provided stable two-axes solutions with stress always < 0.19 , implying a good goodness-of-fit relative to the original sample positions in the multidimensional ordination space. For the full eDNA data matrix, the first vegetation NMDS axis was not related to the first eDNA NMDS axis at all plot sizes (0.01 - 2.02% Deviance Explained; $\text{SES} = 0.07 - 0.14; p > 0.05$), except for the largest $100 \times 100 \text{ cm}^2$ plot size ($\text{SES} = 0.26, p = 0.04$) even though with only 6.53% Deviance Explained. The second vegetation NMDS axis was significantly related to the second eDNA NMDS axis at all but the smallest

plot size, with %Deviance Explained increasing from 18.24% (SES = 0.43; $p < 0.001$; $25 \times 25 \text{ cm}^2$) to 51.81% (SES = 0.72; $p < 0.001$; $100 \times 100 \text{ cm}^2$). For the 0.05% RRA filtered eDNA data matrix, all general linear models at each plot size signaled that both eDNA NMDS axes could consistently and significantly predict their vegetation NMDS counterparts, yet with increasing accuracy as the % explained Deviance increased in absolute terms by 17% (NMDS1) and 21% (NMDS2) from the $10 \times 10 \text{ cm}^2$ to the $100 \times 100 \text{ cm}^2$ plot size. Plant community compositional variation was thus best explained by eDNA cores samples at the $100 \times 100 \text{ cm}^2$ plot size: the first eDNA NMDS axis significantly predicted compositional variation along the first vegetation NMDS axis (SES = 0.73; $p < 0.001$) with 53.84% of the deviance explained. The second eDNA NMDS axis also significantly predicted 68% of the compositional variation along the first vegetation NMDS axis (SES = -0.82; $p < 0.001$). This was also strikingly visible in both NMDS scatterplots (Fig. 6a, b), via comparable clustering patterns of the seven semi-natural grassland sites. For full results on NMDS ordinations, see Table S5 in Supplementary Materials.

Discussion

Our study clearly indicates that soil-based eDNA metabarcoding may indeed deliver on its potential as a stand-alone tool for the discovery of rare plant species and plant community compositional change, aligning with the few studies available in natural terrestrial ecosystems today (Yoccoz et al. 2012, Edwards et al. 2018, Carrasco-Puga et al. 2021, Barnes et al. 2022, Ariza et al. 2023, 2024). Moreover, it is the first study which proves eDNA metabarcoding can detect a single rare species as well as comparatively subtle plant community change in highly diverse semi-natural grassland communities (up to 60 species per m²), which are the most diverse globally at small scales (Wilson et al. 2012), i.e. precisely at those small scales relevant to sampling design. Our findings also convincingly illustrate, however, that sampling design, data curation and analytical parameters need to be specifically tailored to the planned objectives or research questions to ensure that correct ecological or practical inference can reliably be drawn based on the signals picked up in eDNA-generated plant species and community data.

Detection likelihood of rare species

Our results confirm that soil eDNA metabarcoding can repeatedly detect the presence of a single rare plant species amongst 144 recorded grassland species (Carrasco-Puga et al. 2021, but see Hartvig et al. 2021). Even given significant site-specific variation in presence and abundance of the species, robust detection patterns emerged. Proximity to *G. campestris* and population characteristics both shaped the likelihood of detecting *G. campestris* DNA in a soil core (cf. Hartvig et al. 2021). Detection likelihood is indeed expected to decline as populations become smaller and more distant from the point of sampling, mirroring either distance-decaying DNA inputs (via e.g. pollen or propagules (Nakanishi et al. 2021, Nathan & Muller-Landau 2000) and leaf or root tissues (Levy-Booth et al. 2007, Yoccoz et al. 2012)) or other spatially clustered belowground phenomena such as soil seed banks (Plue & Hermy 2012), which *G. campestris* can build (Lennartsson & Oostermeijer 2001). Still, in stark contrast to Hartvig et al. (2021) who only found DNA at distances less than 10 cm from individual orchid plants, traces of *G. campestris* DNA are picked up over significant distances away from the populations, with a 25% detection likelihood at 25m (10% at 50 meters) from the closest individual, likely reflecting a combination of DNA distribution processes such as the long tails typical of seed and pollen dispersal events (Bullock & Clarke 2000, Austerlitz et al. 2004), leaf litter transport (Phillipson et al. 1975), or the persistence of local inter- and extracellular DNA of plants no longer present locally. Regardless of the source, with a more than reasonable effort of only nine soil cores per grassland, eDNA established the species presence in six out of seven grasslands.

However, to be of practical use for further conservation or restoration interventions, the eDNA signal has to be accurately interpreted. In that light, the observation that population density in 2022, the year after soil core sampling, had no significant impact on *G. campestris* detection likelihood seems crucial. This suggests that the eDNA signal is at least dominated by recently shed DNA, hence reflecting the presence of living *G. campestris* individuals. Nevertheless, evidence does suggest that the signal incorporate the

detection of older DNA, originating from e.g. dead or living seeds, partially decomposed plant material or pollen (Carini et al. 2016, Ariza et al. 2023). On the one hand, *G. campestris* DNA was picked up in two out of nine soil cores in a grassland (*Äckelsta*) where the species' presence had not been visually confirmed for over a decade in spite of targeted, annual ocular monitoring efforts. On the other hand, in two grasslands (*Östra Syninge* and *Snedbacken*), population size over the past eight years averaged around one individual annually, suggesting highly limited recent DNA inputs of the small-statured, slender grassland biennial. Consequently, this dual nature of the eDNA signal – combining both contemporary and legacy DNA – implies caution for accurate signal interpretation, yet still offers clear opportunities regarding targeted conservation and restoration efforts.

The discovery of DNA of the species across spatial and temporal gradients using eDNA metabarcoding is an impressive feat, given the species' small-stature, biennial lifespan, low biomass production and low seed production (Lennartsson & Oostermeijer 2001). This realisation is even more impressive given the large amounts of fresh DNA shed by the numerous other grassland species, which co-exist at small spatial scales (Wilson et al. 2012), in highly biologically active semi-natural grassland soils (Siebert et al. 2019), next to methodological repeated subsampling (core sample > 5g soil subsample > 2 µl DNA extract for first PCR > 2 µl PCR product for second PCR) and further technological bottlenecks (e.g. choice of primer or bioinformatic pipeline; sequencing depth) (Deagle et al. 2019, Barnes et al. 2022, Luo et al. 2023), potentially diluting the signal of *G. campestris*. Still, a point of concern is the strong observed reduction in detection likelihood following frequently applied RRA filtering, highlighting the trade-off between (conservative) contamination control and information loss. As noted, *G. campestris* likely only produces limited DNA amounts due to small population sizes, limited above-and belowground biomass production given its small stature and biennial lifespan (Ulrich et al. 2010, Edwards et al. 2018). Thus, filtering out low-abundance reads, while often motivated by concerns about false positives, may disproportionately remove true positives for rare species such as *G. campestris*. Our results therefore clearly caution against applying uniform abundance filtering thresholds in studies or applications using eDNA metabarcoding to target rare species. Instead, filtering strategies should be tailored to objectives: conservative filtering risks undermining the very goals of early detection, red-list monitoring, or restoration planning (Hartvig et al. 2021, Sepp et al. 2021).

Community composition and compositional change

eDNA metabarcoding captured a substantial subset of the extant semi-natural grassland communities. Across all 63 plots, eDNA metabarcoding succeeded to identify ca. 50% of traditionally surveyed plant species, a level of taxonomic overlap closely aligned with earlier reports for soil-based eDNA studies in other natural terrestrial ecosystems (e.g. ~60%, tundra, Edwards et al. 2018; ~50%, desert, Carrasco-Puga et al. 2021; ~60%, coniferous forest, Ariza et al. 2023). This level of consistent compositional overlap across habitat types suggests that the only partial recovery of present species seems a general characteristic of soil eDNA metabarcoding at its current technological state-of-the-art and bottlenecks

(Luo et al. 2023, Barnes et al. 2022), rather than a limitation imposed by the high plant diversity in semi-natural grasslands (Wilson et al. 2012).

Still, given (1) this baseline level of taxonomic overlap, (2) the observed distance-decay detection patterns in *G. campestris* and (3) the ca. 25% probability of failing to detect a species even when sampling in its immediate vicinity (Fig. 2), it is unsurprising that a single soil core did not provide a full account of a plot's plant community across plot sizes, neither in terms of its richness (Fig. 4, 5) nor its composition (Fig. 3b). Moreover, the low, species-richness corrected Raup–Crick similarities suggest that the observed mismatch is not driven by richness differences alone, but by differential species discovery by either method (cf. Carrasco-Puga et al. 2021). Partitioning species richness into its shared, vegetation-only, and eDNA-only cohorts provides valuable insight into this pattern. Across all plot sizes, a significant share of species was detected exclusively by eDNA, highlighting that soil eDNA samples capture species not visible in aboveground surveys at any given plot size. Such detections may include species missed during vegetation inventories, but are more likely to reflect dormant life-cycle stages such as seeds and belowground organs, or persistent intercellular (e.g. pollen, roots) and extracellular DNA from species formerly present or located just outside the sampled plot (cf. Hiiesalu et al. 2012). The number of shared species and eDNA-only species indeed significantly increased and decreased with increasing plot size, respectively, underscoring that numerous species uncovered only by eDNA are species actually present within the local species pool surrounding a sampled plot, reaffirming the highly local nature of an eDNA signal (cf. Hartvig et al. 2021, Ariza et al. 2023). Finally, RRA-filtering magnified species partitioning patterns among cohorts, again asserting this procedure may be overly conservative in removing truthful, and thus ecologically valuable information, as increasingly less species were considered shared or uncovered by eDNA, notwithstanding these species' actual presence. In sum, eDNA analysis of a single soil core does currently not yet function as a direct spatial analogue of aboveground vegetation across plot sizes, but rather as an integrative signal reflecting local species pools across fine spatial grains (cf. Hartvig et al. 2021). Therefore, at plot scales commonly used in semi-natural grassland surveys, composite sampling will be necessary for applications requiring accurate representation of plant community richness and composition, as applied for similar invisible, and highly spatially clustered belowground phenomena (Baker et al. 2009, Muukkonen et al. 2009, Plue & Hermy 2012).

Still, most plant ecological research rarely focuses on knowing the exact number and identity of every species at a given site (Vellend et al. 2008). Instead, we aim to detect changes in species richness and community composition in space and time, helping to pinpoint the environmental drivers which are shaping spatio-temporal plant community responses for conservation purposes. When removing local eDNA signal noise, richness estimates reliably tracked the variation in species richness at the 50 cm × 50 cm and 100 cm × 100 cm plot scales (cf. 40 m × 40 m plots in Barnes et al. 2022). Similarly, NMDS showed that eDNA-derived community composition not only mirrored subtle vegetation-based compositional gradients but quite accurately discriminated individual semi-natural grasslands (cf. Hartvig et al. 2021, Ariza et al. 2024). Combined, this leads to the conclusion that although eDNA underestimates

absolute species richness by capturing only a subset of present species in individual soil cores at all plot sizes, this species subset both retains sufficient plant community structure to distinguish plots of low versus high diversity as well as identify the core ecological gradients shaping the spatial and environmental structure of semi-natural grassland communities, both within and across sampled semi-grasslands. This aligns with Vellend et al. (2008)’s argument demonstrating that subsets of species, even if incomplete (missing in some cases up to 80% of occurring species), can still reliably reflect patterns of variation in richness and composition if they include sufficient information from the broader community, i.e. if subsets are non-random and representative. As the Raup–Crick similarities confirm eDNA species assemblage to be a non-random, compositional subset of the observed community, the strong synchrony in conveying similar ecological information about spatial patterns of plant community turnover between eDNA and classical inventories therefore strongly suggests that eDNA indeed uncovers representative – and thus ecologically meaningful – subsets of species.

Conclusion

The purpose of ecological surveys is to learn as much as we can with limited resources when, for example, mapping rare species distributions or quantifying changes in species richness or community composition by means of efficient, representative sampling strategies (cf. Vellend et al. 2008; Plue & Hermy 2012). eDNA metabarcoding is the latest addition to this sampling toolbox, with results favourably arguing for soil-based eDNA possibly becoming a successful stand alone tool at delivering a cost- and resource-effective means for surveying diverse terrestrial plant communities, pending that study objectives, sampling strategies and analytical choices are well-aligned. Our results highlight both the main practical consequences for sampling strategies and analytical pathways which arise when choosing between a focus on detailed, complete plant community inventories including rare species detection versus unveiling spatial and temporal patterns in plant communities. Therefore, we believe that soil-based eDNA metabarcoding may presently not always replace traditional surveys for exhaustive plot-level species inventories, yet it is already a potent tool in providing a scalable, observer-independent tool for detecting biodiversity patterns, monitoring compositional change, and thus effectively informing conservation and restoration efforts.

References

- Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.
- Anonymous (2026) Data from: Aligning terrestrial eDNA sampling and analytical choices to effectively capture both rare species and compositional variation in grassland plant communities. Dryad.
- Ariza, M., Fouks, B., Mauvisseau, Q., Halvorsen, R., Alsos, I. G., & de Boer, H. J. (2023). Plant biodiversity assessment through soil eDNA reflects temporal and local diversity. *Methods in Ecology and Evolution*, 14(2), 415-430.
- Ariza, M., Engelstad, M., Lieungh, E., Laux, M., Ready, J., Mauvisseau, Q., ... & de Boer, H. J. (2024). Evaluating the feasibility of using plant-specific metabarcoding to assess forest types from soil eDNA. *Applied Vegetation Science*, 27(4), e12806.
- Auffret, A. G., Kimberley, A., Plue, J., & Waldén, E. (2018). Super-regional land-use change and effects on the grassland specialist flora. *Nature Communications*, 9(1), 3464.
- Austerlitz, F., Dick, C. W., Dutech, C., Klein, E. K., Oddou-Muratorio, S., Smouse, P. E., & Sork, V. L. (2004). Using genetic markers to estimate the pollen dispersal curve. *Molecular Ecology*, 13(4), 937-954.
- Azevedo, L. B., van Zelm, R., Hendriks, A. J., Bobbink, R., & Huijbregts, M. A. (2013). Global assessment of the effects of terrestrial acidification on plant species richness. *Environmental Pollution*, 174, 10-15.
- Bachman, S. P., Field, R., Reader, T., Raimondo, D., Donaldson, J., Schatz, G. E., & Nic Lughadha, E. (2019). Progress, challenges and opportunities for Red Listing. *Biological Conservation*, 234, 45-55.
- Baker, K. L., Langenheder, S., Nicol, G. W., Ricketts, D., Killham, K., Campbell, C. D., & Prosser, J. I. (2009). Environmental and spatial characterisation of bacterial community composition in soil to inform sampling strategies. *Soil Biology and Biochemistry*, 41(11), 2292-2298.
- Banks-Leite, C., Ewers, R. M., Folkard-Tapp, H., & Fraser, A. (2020). Countering the effects of habitat loss, fragmentation, and degradation through habitat restoration. *One Earth*, 3(6), 672-676.
- Barnes, C. J., Nielsen, I. B., Aagaard, A., Ejrnæs, R., Hansen, A. J., & Frøslev, T. G. (2022). Metabarcoding of soil environmental DNA replicates plant community variation but not specificity. *Environmental DNA*, 4(4), 732-746.
- Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*, 115(25), 6506-6511.
- Bullock, J. M., & Clarke, R. T. (2000). Long distance seed dispersal by wind: measuring and modelling the tail of the curve. *Oecologia*, 124(4), 506-521.
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13, 581-583.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1), 421.
- Carini, P., Marsden, P. J., Leff, J. W., Morgan, E. E., Strickland, M. S., & Fierer, N. (2016). Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nature Microbiology*, 2(3), 1-6.

621 Carrasco-Puga, G., Díaz, F. P., Soto, D. C., Hernández-Castro, C., Contreras-López, O., Maldonado, A.,
622 ... & Gutiérrez, R. A. (2021). Revealing hidden plant diversity in arid environments. *Ecography*, 44(1), 98-
623 111.

624 Catford, J.A., Bode, M. & Tilman, D. Introduced species that overcome life history trade-offs can cause
625 native extinctions. *Nature Communications*, 9, 2131 (2018).

626 Ceulemans, T., Stevens, C.J., Duchateau, L., Jacquemyn, H., Gowing, D.J.G., Merckx, R., Wallace, H., van
627 Rooijen, N., Goethem, T., Bobbink, R., Dorland, E., Gaudnik, C., Alard, D., Crockett, E., Muller, S.,
628 Dise, N.B., Dupré, C., Diekmann, M. and Honnay, O. (2014), Soil phosphorus constrains biodiversity
629 across European grasslands. *Global Change Biology*, 20: 3814-3822.

630 Chen, X., Chen, H. Y., Chen, C., Ma, Z., Searle, E. B., Yu, Z., & Huang, Z. (2020). Effects of plant
631 diversity on soil carbon in diverse ecosystems: A global meta-analysis. *Biological Reviews*, 95(1), 167-183.

632 Cousins, S.A.O., Auffret, A. G., Lindgren, J., & Tränk, L. (2015). Regional-scale land-cover change during
633 the 20th century and its consequences for biodiversity. *Ambio*, 44, 17-27.

634 Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., ... & Eveson, J.
635 P. (2019). Counting with DNA in metabarcoding studies: How should we convert sequence reads to
636 dietary data? *Molecular ecology*, 28(2), 391-406.

637 de Vries, J., Heine, M., Harms, K., & Wackernagel, W. (2003). Spread of recombinant DNA by roots and
638 pollen of transgenic potato plants, identified by highly specific biomonitoring using natural
639 transformation of an *Acinetobacter* sp. *Applied and Environmental Microbiology*, 69, 4455-4462.

640 Duley, E., Iribar, A., Bisson, C., Chave, J., & Donald, J. (2023). Soil environmental DNA metabarcoding
641 can quantify local plant diversity for biomonitoring across varied environments. *Restoration Ecology*,
642 31(4), e13831.

643 Edwards, M. E., Alsos, I. G., Yoccoz, N., Coissac, E., Goslar, T., Gielly, L., ... & Taberlet, P. (2018).
644 Metabarcoding of modern soil DNA gives a highly local vegetation signal in Svalbard tundra. *The*
645 *Holocene*, 28(12), 2006-2016.

646 Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance?
647 Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol.
648 *PloS one*, 10(7), e0130324.

649 Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., &
650 Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature*
651 *Biotechnology*, 38, 276–278

652 Frøslev, T. G., Ejrnæs, R., Hansen, A. J., Bruun, H. H., Nielsen, I. B., Ekelund, F., ... & Kjøller, R. (2023).
653 Treated like dirt: Robust forensic and ecological inferences from soil eDNA after challenging sample
654 storage. *Environmental DNA*, 5(1), 158-174.

655 Grueber, C.E., Nakagawa, S., Laws, R.J., & Jamieson, I.G. (2011). Multimodel inference in ecology and
656 evolution: challenges and solutions. *Journal of Evolutionary Biology*, 24, 699-711.

657 Haddad, N. M., Crutsinger, G. M., Gross, K., Haarstad, J., & Tilman, D. (2011). Plant diversity and the

658 stability of food webs. *Ecology letters*, 14(1), 42-46.

659 Hartvig, I., Kosawang, C., Kjær, E. D., & Nielsen, L. R. (2021). Detecting rare terrestrial orchids and
660 associated plant communities from soil samples with eDNA methods. *Biodiversity and Conservation*,
661 30(13), 3879-3901.

662 Heckenhauer, J., Barfuss M. H. J., Samuel R. (2016) Universal multiplexable matK primers for DNA
663 barcoding of angiosperms. *Appl. Plant Sci.* 8;4(6):apps.1500137. doi: 10.3732/apps.1500137

664 Hiiesalu, I., Oepik, M., Metsis, M., Lilje, L., Davison, J., Vasar, M., ... & Pärtel, M. (2012). Plant species
665 richness belowground: higher richness and new patterns revealed by next-generation sequencing.
666 *Molecular Ecology*, 21(8), 2004-2016.

667 IUCN. 2022. The IUCN Red List of Threatened Species. Version 2022-2. <https://www.iucnredlist.org>.

668 Kier, G., Mutke, J., Dinerstein, E., Ricketts, T. H., Küper, W., Kreft, H., & Barthlott, W. (2005). Global
669 patterns of plant diversity and floristic knowledge. *Journal of Biogeography*, 32(7), 1107-1116.

670 Kimmel, K., Clark, M., & Tilman, D. (2022). Impact of multiple small and persistent threats on extinction
671 risk. *Conservation Biology*, 36(5), e13901.

672 Koleff, P., Gaston, K. J., & Lennon, J. J. (2003). Measuring beta diversity for presence-absence data.
673 *Journal of Animal Ecology*, 72(3), 367-382.

674 Lennartsson, T., & Oostermeijer, J. G. B. (2001). Demographic variation and population viability in
675 *Gentianella campestris*: effects of grassland management and environmental stochasticity. *Journal of*
676 *Ecology*, 451-463.

677 Letourneau, D. K., Armbrrecht, I., Rivera, B. S., Lerma, J. M., Carmona, E. J., Daza, M. C., ... & Trujillo,
678 A. R. (2011). Does plant diversity benefit agroecosystems? A synthetic review. *Ecological Applications*,
679 21(1), 9-21.

680 Levy-Booth, D. J., Campbell, R. G., Gulden, R. H., Hart, M. M., Powell, J. R., Klironomos, J. N., ... &
681 Dunfield, K. E. (2007). Cycling of extracellular DNA in the soil environment. *Soil Biology and*
682 *Biochemistry*, 39(12), 2977-2991.

683 Luo, M., Ji, Y., Warton, D., & Yu, D. W. (2023). Extracting abundance information from DNA-based
684 data. *Molecular Ecology Resources*, 23(1), 174-189.

685 Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads.
686 *EMBnet. journal*, 17(1), 10-12.

687 Montràs-Janer, T., Suggitt, A. J., Fox, R., Jönsson, M., Martay, B., Roy, D. B., ... & Auffret, A. G. (2024).
688 Anthropogenic climate and land-use change drive short-and long-term biodiversity shifts across taxa.
689 *Nature Ecology & Evolution*, 1-13.

690 Muukkonen, P., Häkkinen, M., & Mäkipää, R. (2009). Spatial variation in soil carbon in the organic layer
691 of managed boreal forest soil—implications for sampling design. *Environmental Monitoring and*
692 *Assessment*, 158, 67-76.

693 Nakanishi, A., Goto, S., Sumiyoshi, C., & Isagi, Y. (2021). Directional seed and pollen dispersal and their
694 separate effects on anisotropy of fine-scale spatial genetic structure among seedlings in a dioecious, wind-

695 pollinated, and wind-dispersed tree species, *Cercidiphyllum japonicum*. Ecology and Evolution, 11(12), 7754-
696 7767.

697 Nathan, R., & Muller-Landau, H. C. (2000). Spatial patterns of seed dispersal, their determinants and
698 consequences for recruitment. Trends in Ecology & Evolution, 15(7), 278-285.

699 Nic Lughadha, E., Bachman, S. P., Leão, T. C., Forest, F., Halley, J. M., Moat, J., ... & Walker, B. E.
700 (2020). Extinction risk and threats to plants and fungi. Plants, People, Planet, 2(5), 389-408.

701 Pardoe, H. S. (1996). Micro-scale patterns of modern pollen deposition within three alpine plant
702 communities. New Phytologist, 132(2), 327-341.

703 Phillipson, J., Putman, R. J., Steel, J., & Woodell, S. R. J. (1975). Litter input, litter decomposition and the
704 evolution of carbon dioxide in a beech woodland—Wytham Woods, Oxford. Oecologia, 20(3), 203-217.

705 Plue, J., & Cousins, S. A. (2013). Temporal dispersal in fragmented landscapes. Biological Conservation,
706 160, 250-262.

707 Plue, J., & Hermy, M. (2012). Consistent seed bank spatial structure across semi-natural habitats
708 determines plot sampling. Journal of Vegetation Science, 23(3), 505-516.

709 Poté, J., Rosselli, W., Wigger, A., & Wildi, W. (2007). Release and leaching of plant DNA in unsaturated
710 soil columns. Ecotoxicology and Environmental Safety, 68(2), 293-298.

711 Pärtel, M., & Wilson, S. D. (2002). Root dynamics and spatial pattern in prairie and forest. Ecology, 83(5),
712 1199-1203.

713 Rodrigues, A. S., Pilgrim, J. D., Lamoreux, J. F., Hoffmann, M., & Brooks, T. M. (2006). The value of the
714 IUCN Red List for conservation. Trends in Ecology & Evolution, 21(2), 71-76.

715 Román-Palacios, Cristian, and John J. Wiens. "Recent responses to climate change reveal the drivers of
716 species extinction and survival." Proceedings of the National Academy of Sciences 117.8 (2020): 4211-
717 4217.

718 Sandor, Manette E., Chris S. Elphick, and Morgan W. Tingley. Extinction of biotic interactions due to
719 habitat loss could accelerate the current biodiversity crisis. Ecological Applications 32.6 (2022): e2608.

720 Sepp, S. K., Davison, J., Moora, M., Neuenkamp, L., Oja, J., Roslin, T., ... & Zobel, M. (2021). Woody
721 encroachment in grassland elicits complex changes in the functional structure of above-and belowground
722 biota. Ecosphere, 12(5), e03512.

723 Snyder (2011) Leaving home ain't easy: non-local seed dispersal is only evolutionarily stable in highly
724 unpredictable environments. Proc. R. Soc. B.278739–744

725 Strzepek, K., Yohe, G., Neumann, J., & Boehlert, B. (2010). Characterizing changes in drought risk for
726 the United States from climate change. Environmental Research Letters, 5(4), 044012.

727 Ståhl, G., Allard, A., Esseen, P. A., Glimskär, A., Ringvall, A., Svensson, J., ... & Inghe, O. (2011).
728 National Inventory of Landscapes in Sweden (NILS)—scope, design, and experiences from establishing a
729 multiscale biodiversity monitoring system. Environmental monitoring and assessment, 173, 579-595.

730 Ulrich, W., Ollik, M., & Ugland, K. I. (2010). A meta-analysis of species–abundance distributions. Oikos,
731 119(7), 1149-1155.

732 Ulrich, W., Olszewski, P., Puchalka, R., Sewerniak, P., & Koprowski, M. (2021). Inter-and intraspecific
733 spatial distributions, spatial segregation by dominants and emergent neutrality in understorey plants.
734 *Oikos*, 130(10), 1813-1822.

735 Vellend, M., Verheyen, K., Flinn, K. M., Jacquemyn, H., Kolb, A., Van Calster, H., ... & Hermy, M.
736 (2007). Homogenization of forest plant communities and weakening of species–environment
737 relationships via agricultural land use. *Journal of Ecology*, 95(3), 565-573.

738 Vellend, M., Lilley, P. L., & Starzomski, B. M. (2008). Using subsets of species in biodiversity
739 surveys. *Journal of Applied Ecology*, 45(1), 161-169.

740 Vellend, M., Baeten, L., Myers-Smith, I. H., Elmendorf, S. C., Beauséjour, R., Brown, C. D., ... & Wipf, S.
741 (2013). Global meta-analysis reveals no net change in local-scale plant biodiversity over time. *Proceedings*
742 *of the National Academy of Sciences*, 110(48), 19456-19459.

743 Verheyen, K., Bažány, M., Češko, E., Chudomelová, M., Closset-Kopp, D., Czortek, P., ... & Baeten, L.
744 (2018). Observer and relocation errors matter in resurveys of historical vegetation plots. *Journal of*
745 *Vegetation Science*, 29(5), 812-823.

746 Wan, N. F., Zheng, X. R., Fu, L. W., Kier, L. P., Zhang, Z., Chaplin-Kramer, R., ... & Li, B. (2020).
747 Global synthesis of effects of plant species diversity on trophic groups and interactions. *Nature Plants*,
748 6(5), 503-510.

749 Watson, J. E., Jones, K. R., Fuller, R. A., Marco, M. D., Segan, D. B., Butchart, S. H., ... & Venter, O.
750 (2016). Persistent disparities between recent rates of habitat conversion and protection and implications
751 for future global conservation targets. *Conservation Letters*, 9(6), 413-421.

752 Wilson, J. B., Peet, R. K., Dengler, J., & Pärtel, M. (2012). Plant species richness: the world records.
753 *Journal of vegetation Science*, 23(4), 796-802.

754 Yoccoz, N. G., Bråthen, K. A., Gielly, L., Haile, J., Edwards, M. E., Goslar, T., ... & Taberlet, P. (2012).
755 DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular ecology*, 21(15), 3647-
756 3655.

757 Zhang, X., Wan, H., Zwiers, F. W., Hegerl, G. C., & Min, S. K. (2013). Attributing intensification of
758 precipitation extremes to human influence. *Geophysical Research Letters*, 40(19), 5252-5257

759 Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., ... & Taberlet, P. (2019). DNA
760 metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular*
761 *ecology*, 28(8), 1857-1862.

762 Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). Mixed effects models and
763 extensions in ecology with R (Vol. 574, p. 574). New York: Springer.

Figures

Fig. 1 eDNA sampling design applied across 7 managed semi-natural grasslands in Stockholm county, Sweden (a). Nine eDNA soil core samples (gold arrows) were collected per grassland (b), at increasing distances away from the centre of occurrence of *Gentianella campestris* (purple polygon) in the grassland (green polygon). Around each core sample, plant species presence was recorded in four plots of increasing sizes (10, 25, 50 and 100 cm) centred around the soil core (c). The sampling design aimed to investigate terrestrial eDNA detection likelihood of a rare grassland plant specialist and eDNA's potential to representatively capture variation in plant community composition.

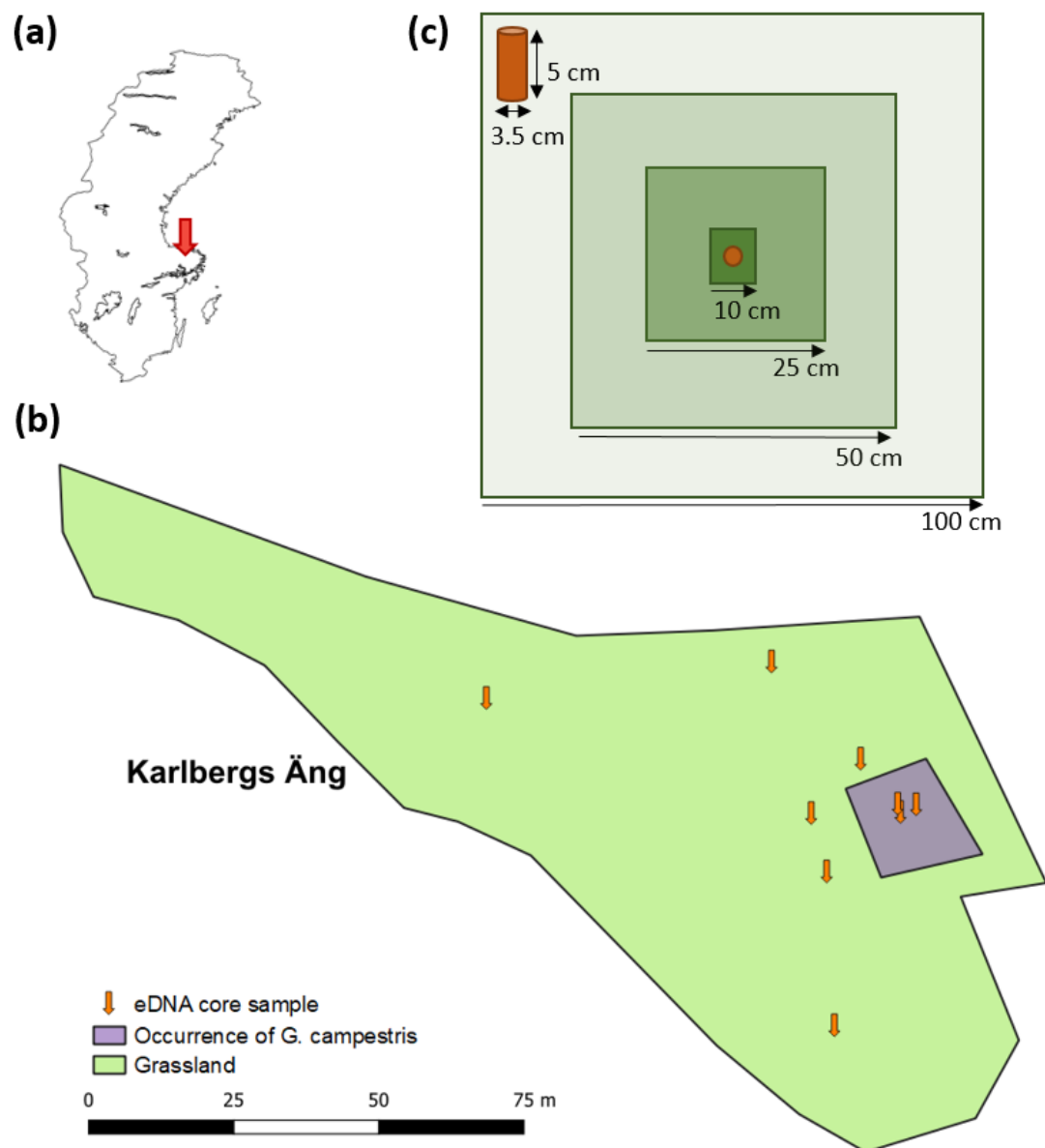


Fig. 2 Direct effects of distance to the nearest individual (a, e), population size (b, f) and population density in 2021 (c, g) and 2022 (d, h) on the detection likelihood of the grassland plant specialist *Gentianella campestris* ($p(G. campestris)$) across 63 soil core samples, both for the full presence/absence data (top row, gold) and the RRA filtered presence/absence data (bottom row, turquoise). Standardized effect sizes (SES) and associated p-values are calculated via mixed effects logistic regressions using scaled predictor variables to enable comparison of the strength of the various predictor effects across models. Subplots (a) - (d) are ordered from right to left in decreasing importance of the predictors effect size on *G. campestris* detection likelihood.

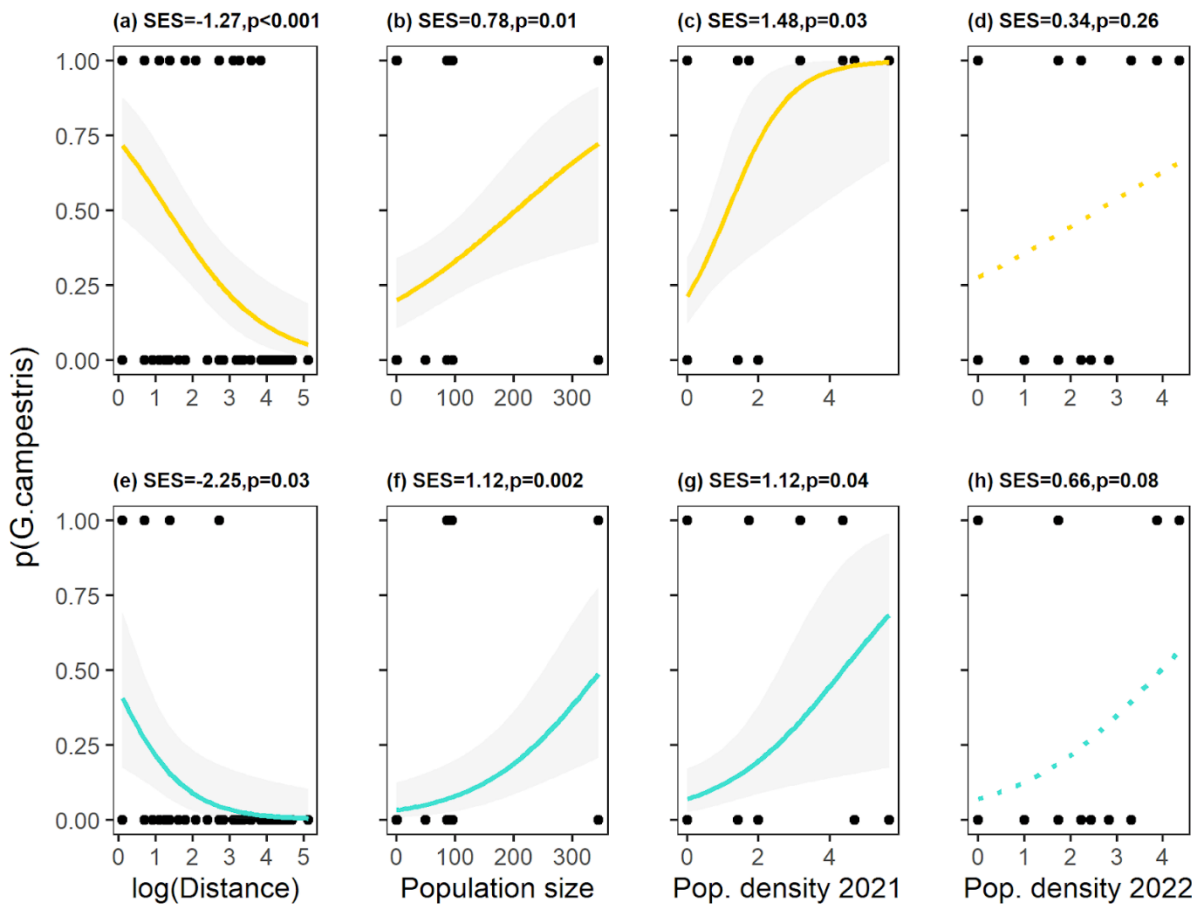
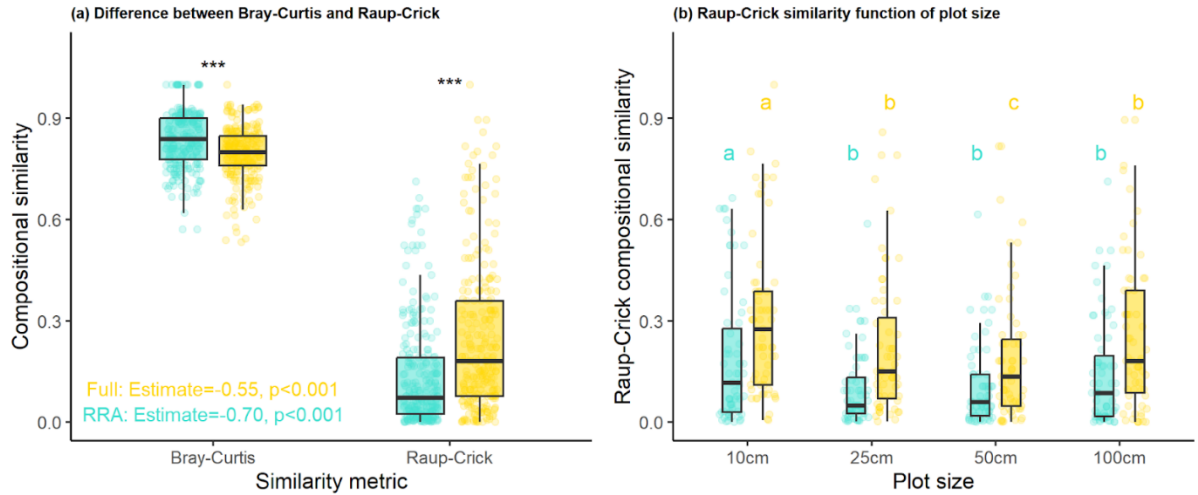


Fig. 3 The impact of choice of similarity metric (a), plant inventory plot size (b) and data curation ((a) and (b)) on the compositional similarity between plant communities assessed via eDNA and traditional plot inventories. Colour coding visualizes the effect of data curation, i.e. when compositional similarity was calculated using either the full (gold) or 0.05% RRA-filtered eDNA dataset (turquoise).



(a) Bray-Curtis similarity is significantly larger than Raup-Crick similarity, both when using the full (gold) and 0.05% RRA-filtered eDNA dataset (turquoise) to compare compositional similarity with the extant plant community. RRA filtering lead to significantly higher and lower similarity in the Bray-Curtis and Raup-Crick metric, respectively (***, $p < 0.001$; paired t-test). (b) Homogeneous subsets

Fig. 4 Partitioning of plant species uncovered by both eDNA and plant inventories at each sampling point as a function of plot size, before (a) and after (b) data curation by applying a 0.05% RRA filter on the full read abundance eDNA data. Species are partitioned as being uncovered: by the plant inventory only (red), by the eDNA assay only (green) or by both plant inventory and eDNA assay (blue). Homogeneous subsets identify significant shifts in species partitioning between inventory type with plot size.

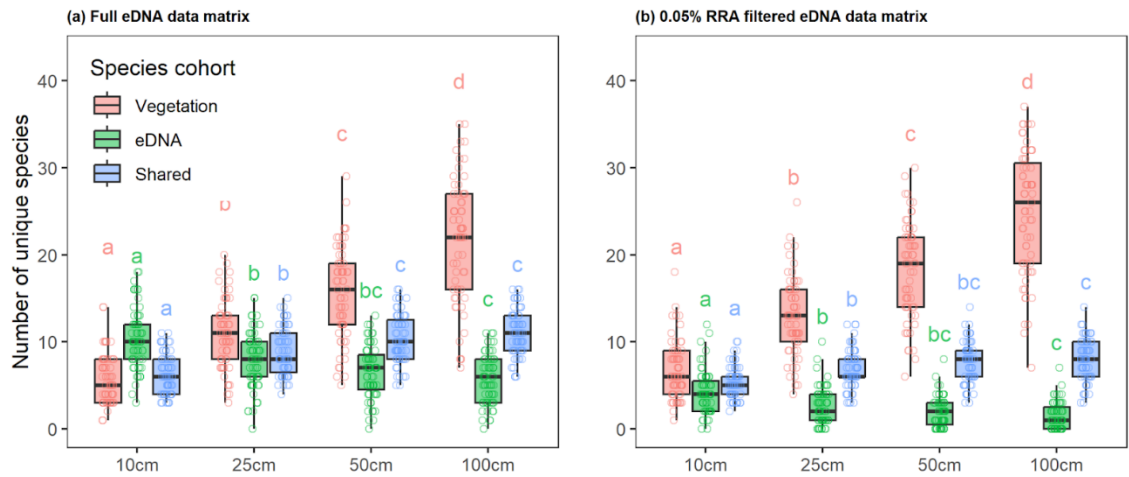


Fig. 5 Relationships between species richness of the plant community identified using eDNA metabarcoding of 63 individual soil cores and quantified using plot-based inventories of different size (10 (a, e), 25 (b, f), 50 (c, g) and 100 cm (d, h)) nested within one another around each soil core. Relationships are presented for both the full presence/absence data (top row, gold) and the 0.05% RRA filtered presence/absence data (bottom row, turquoise), with full regression lines marking significant relationships at $p < 0.05$. Standardized effect sizes (SES) and associated p -values are calculated via general linear models using scaled dependent variables to enable comparison of effect sizes across models.

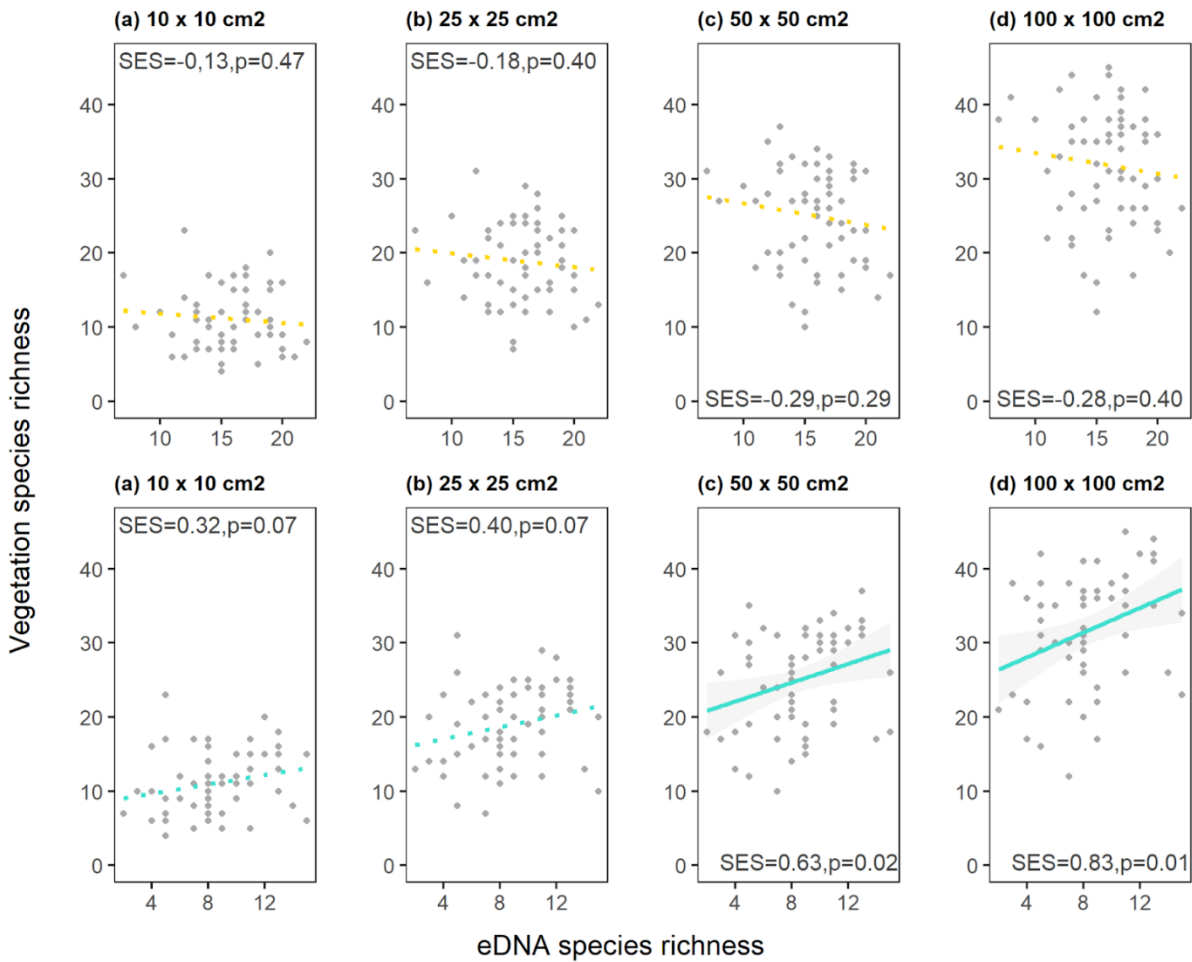
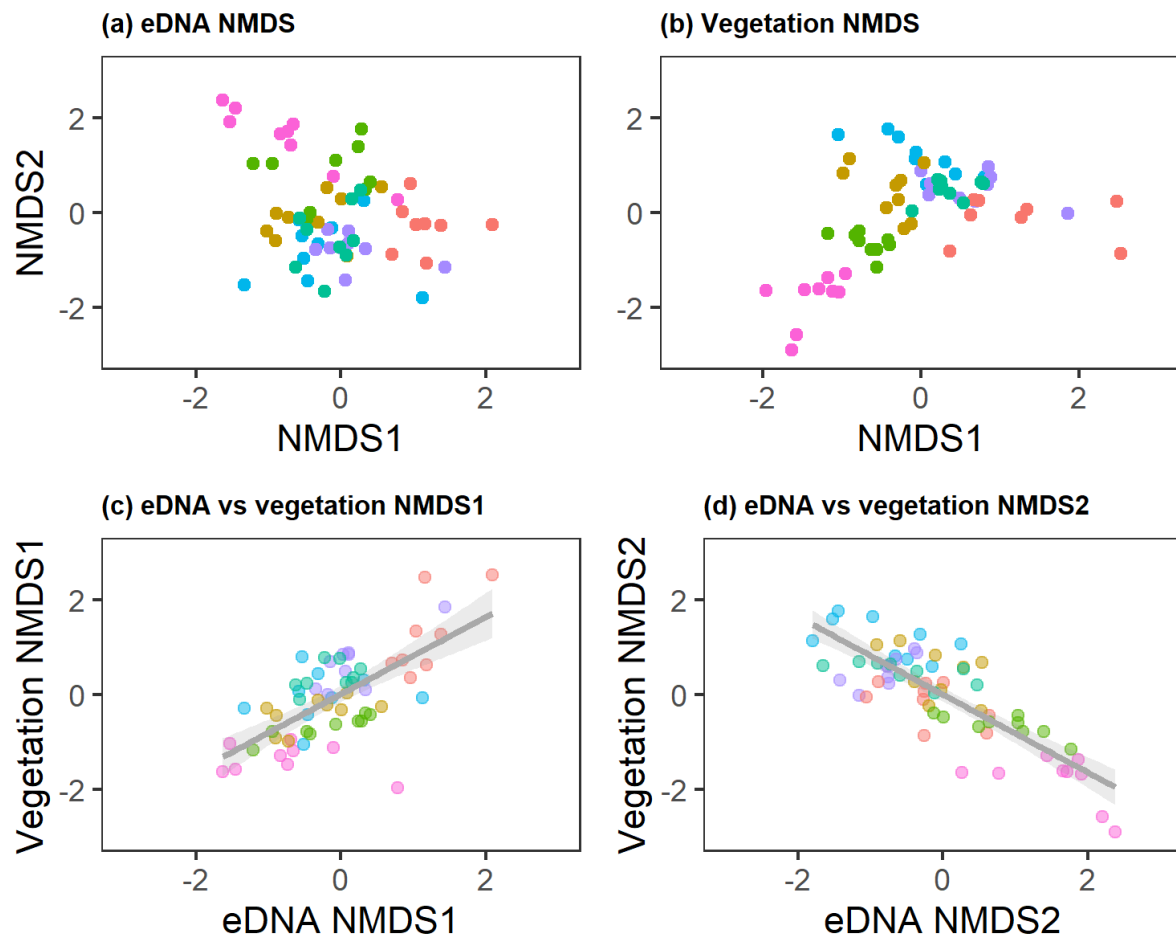


Fig. 6 NMDS ordinations (based on Raup-Crick similarities) run separately on the 0.05% RRA filtered eDNA species $\times$ sample matrix (a) and observed extant plant communities within 100×100 cm² plots centered around each of the 63 eDNA soil samples (b). Color coding identifies each of the seven semi-natural grasslands included in the study. Regression lines in subplots (c) and (d) highlight significant relations between community compositional gradients (NMDS axes) discovered separately in the eDNA and observed plant community data matrices.



Supplementary Information

These Supplementary Materials include:

Tables S1, S2, S3, S4 & S5

Legends for Datasets S1 to S3

Datasets S1 to S3

Table S1 Basic characteristics of the seven semi-natural grasslands in Stockholm county (Sweden) in which 9 soil cores were collected for eDNA metabarcoding to identify .

SiteID	X	Y	Grassland area [m2]	Management	<i>G. campestris</i> population size 2013-2021
Bygget	59.79	18.25	42130	Grazing	344
Karlberg äng	59.82	18.24	15430	Mowing	96
Målsjön	59.79	18.25	12720	Grazing	49
Åckelsta	59.69	18.06	14050	Grazing	0
Snedbacken	59.96	18.07	29390	Grazing	1
Svartkärret	59.72	18.33	5050	Mowing	86
Östra Syninge	59.75	18.51	50425	Grazing	1

Table S2 PCR primers used to amplify the matK region. Primer sequences were modified from Heckenhauer et al. (2016) to include a M13 sequence for the dual indexes to ligate to and a 5' Amino Modifier C6 modification (synthesised by Integrated DNA Technologies inc., IDT).

M13_matK-413f-1	/5AmMC6/ GTAAAACGACGGCCAGTTAATTTACRATCAATTCATTCAATATTTCC
M13_matK-413f-2	/5AmMC6/ GTAAAACGACGGCCAGTTAATTTACGATCYATTTCATTCAATATTTCC
M13_matK-413f-3	/5AmMC6/ GTAAAACGACGGCCAGTTAATTTACGATCAATTCATTCAACATTTCC
M13_matK-413f-4	/5AmMC6/ GTAAAACGACGGCCAGTTAATTTMCRATCAATTCATTCCATATTTCC
M13_matK-413f-5	/5AmMC6/ GTAAAACGACGGCCAGTTAATTTACGATCAATTCATTCACTATTTCC
M13_matK-1227r-1	/5AmMC6/ CAGGAAACAGCTATGACGARGAYCCRCRTRRATAATGAGAAAGATTT
M13_matK-1227r-2	/5AmMC6/ CAGGAAACAGCTATGACGAAGAYCCGCTATGATAATGAGAAAGGTTT
M13_matK-1227r-3	/5AmMC6/ CAGGAAACAGCTATGACGARGATCCRCRTRRATAATGAAAAAGATTT
M13_matK-1227r-4	/5AmMC6/ CAGGAAACAGCTATGACGARGATCCRCRTRRATAATGAGAAAAATTT
M13_matK-1227r-5	/5AmMC6/ CAGGAAACAGCTATGACGARGATCCRCRTRRATAATGAGAAATATTT

Table S3 Parameters used in the bioinformatics pipeline used for taxonomic assignment of ASVs using the dada2 R package (Callahan et al. 2016).

```
nextflow run nf-core/ampliseq \
  -resume \
  -r 2.7.0 \
  -profile singularity \
  --input "Gentianella_campestris_matK_samples.tsv" \
  --metadata "Gentianella_metadata.tsv" \
  --pachio \
  --min_len 700 \
  --skip_qiime \
  --sample_inference pseudo \
  --FW_primer TAATTTMCRATCHATTTCMMYATTTCC \
  --RV_primer GARGAYCCRCRTRATAATGARAAADRITT \
  --dada_ref_tax_custom Swedish_plants_matK.v0.14xx.COMB_SP_NAMES.fasta \
  --dada_ref_tax_custom_sp SP_Swedish_plants_matK.v0.14xx.COMB_SP_NAMES.fasta \
  --dada_assign_taxlevels "Domain,Phylum,Class,Order,Family,Genus,Species" \
  --outdir "./result" \
  --seed 999 \
  --max_memory "30 GB" \
  --max_cpus 4
```

Table S4 Detection likelihood of *Gentianella campestris* in an eDNA soil core sample in function of *G. campestris* population and sampling characteristics, modelled using logistic regression models, all containing *site* as a random model term. Base models contained a single scaled explanatory variable (returning Standardized Effect Sizes (SES) for comparability between models), using likelihood ratio tests to assess model improvement to a null model. In the additive models, we investigated the importance of *Distance* (1) and *2021 Population density* (2) by adding these variables separately to the base model containing *Population size* as explanatory variable, using likelihood ratio test to assess model improvement. This analysis was done both on the presence of *G. campestris* in an eDNA soil core sample before (full) and after (RRA) data curation, i.e. applying a <0.05% RRA omission filter on the raw ASV read abundance matrix. Distance was log-transformed whereas Population densities were square-root transformed for normality, prior to scaling.

p(<i>G. campestris</i>)			Full			RRA		
n = 63			Base	Additive¹	Additive²	Base	Additive¹	Additive²
Distance	Model	SES	-1.27	-1.16		-2.25	-1.63	
		p	0.001	0.002		0.03	0.03	
	Likelihood	Chi²	14.00	12.75		10.73	8.08	
		p	<0.001	<0.001		0.001	0.005	
Population size	Model	SES	0.78	0.73	0.60	1.12	1.19	1.10
		p	0.01	0.03	0.06	0.002	0.008	0.005
	Likelihood	Chi²	5.21	3.96	3.25	7.53	4.88	5.00
		p	0.02	0.046	0.07	0.006	0.03	0.03
Population density - 2021	Model	SES	1.48		1.20	1.12		0.65
		p	0.03		0.02	0.04		0.046
	Likelihood	Chi²	12.37		10.42	6.16		3.64
		p	<0.001		0.001	0.01		0.06
Population density - 2022	Model	SES	0.34			0.66		
		p	0.26			0.08		
	Likelihood	Chi²	1.29			3.12		
		p	0.25			0.08		

942
943
944
945
946
947
948
949
950

Dataset S1 Soil core and *Gentianella campestris* population characteristics

For each of 63 eDNA soil cores collected, this dataset lists: (1) the semi-natural grassland and (2) its size [m²] in which soil cores were collected, (3) the latitude and longitude per individual core, (4) the grassland management, (5) the distance of the soil core to the nearest and last known observation of *Gentianella campestris*, (6) 2021 and (7) 2022 *G. campestris* population density, i.e. the number of *G. campestris* individuals counted in a 2m radius surrounding each soil core, (8) mean population size of *G. campestris* in the last 8 years, (9) the last year *G. campestris* was observed in the semi-natural grassland and finally, if *G. campestris* was detected by eDNA [0/1] before (10) and after (11) a 0.05% RRA filter was applied to the raw ASV read eDNA plant community matrix (Dataset S3).

Dataset S2 Abundance-based plant community matrix

Full 165 species × 252 plot-based plant community matrix. Four nested plant relevés of different plot sizes (10, 25, 50 and 100 cm) were performed centred around each of 63 individual soil cores collected for eDNA analysis, at a rate of 9 soil cores in each of 7 semi-natural grasslands in Stockholm county, Sweden.

Dataset S3 ASV-based eDNA plant community matrix

Full 72 species × 63 eDNA-based plant community matrix, after manual curation of the raw ASV abundance matrix. Values represent the number of ASV reads assigned to a species via taxonomic bioinformatic pipelines (against libraries consisting of the species present in the Swedish region of Uppland) and further manual data curation via BLASTn.