A systematic survey of regional multitaxon biodiversity: evaluating strategies and coverage

Ane Kirstine Brunbjerg[1], Hans Henrik Bruun[2], Lars Brøndum[3], Aimée T. Classen[4,5], Kåre Fog[6], Tobias G. Frøslev[7], Irina Goldberg[2], Morten D.D. Hansen[3], Toke T. Høye[1,8*], Thomas Læssøe[2], Gregory S. Newman[4,5], Lars Skipper[3], Ulrik Søchting[2], Rasmus Ejrnæs[1]

[1] *Section for Biodiversity & Conservation, Department of Bioscience, Aarhus University, DK-8410 Rønde, Denmark (present address)*

[2] *Section for Ecology and Evolution, Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark / Natural History Museum of Denmark, Universitetsparken 15, University of Copenhagen, DK-2100 Copenhagen, Denmark*

[3] *Natural History Museum Aarhus, DK-8000 Aarhus C, Denmark*

[4] *Rubenstein School of Environment and Natural Resources, University of Vermont, Burlington, VT 05405, USA,*

[5] *Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, DK-1350 Copenhagen, Denmark.*

[6] *Hesselholm 107, 3670 Veksø, Denmark*

[7] *Center for GeoGenetics, University of Copenhagen, DK-2100 Copenhagen, Denmark*

[8] *Arctic Research Centre, Aarhus University, Ny Munkegade 114, building 1540, DK-8000 Aarhus C, Denmark*

* Corresponding author: Tel.: +45 87158892, E-mail address: tth@bios.au.dk

Abstract

Background: In light of the biodiversity crisis and our limited ability to explain variation in biodiversity, it is time to rethink the way we study biodiversity and its causes. Inspired by the recently published ecospace framework, we developed a protocol for environmental and biotic mapping that is scalable to habitats, ecosystems and biomes. We applied our protocol as part of a comprehensive biodiversity study in Denmark. We selected study sites (40 × 40 m) using stratified random sampling along the major environmental gradients underlying biotic variation. Using standard methods, we collected vascular plant, bryophyte, macrofungi, lichen, gastropod and arthropod species lists for each site. To evaluate sampling efficiency, we calculated regional coverage (relative to the number of species known from Denmark per taxonomic group), and project scale coverage (i.e., based on the sample coverage per taxonomic groups). To cover eukaryotic organisms that are less easily targeted by classical inventories (e.g., nematodes, "non-fruiting" fungi) we collected soil samples for environmental DNA analyses. Finally, to assess site conditions, we conducted a comprehensive mapping of abiotic conditions (position), biotic resources (expansion of organic carbon) and habitat continuity (spatial and temporal).

Results: The 130 study sites covered 0.0005% of the Danish terrestrial area (~42,500 km^2). We found 2040 species of macrofungi (62% of the Danish fungal pool), 663 vascular plant species (42%), 254 bryophyte species (41%) and 202 lichen species (20%). For invertebrates, we observed 334 spider species (59%), 126 carabid beetle species (38%) and 105 hoverfly species (36%). Overall, sample coverage was high across taxonomic groups, indicating that 130 sites were sufficient to represent the variation in biodiversity across Denmark. This inventory is unprecedented in detail and resulted in the discovery of 150 species with no previous record for

Denmark. Comparison of soil DNA with observed plants was both strong and confirmative for a recovery of plant biota by soil-derived DNA.

Conclusions: We successfully covered the majority of targeted biodiversity across Denmark using an approach that includes habitat coverage, multi-taxon biodiversity assessment, and ecospace mapping. Our approach can be readily applied to assess biodiversity for other ecoregions.

Keywords: Abiotic gradients, Biotic factors, Continuity, Denmark, Disturbance, eDNA, Moisture, Productivity

Background

The vast number of species on Earth have yet to be described, challenging our understanding of biodiversity [1]. For a deeper understanding of what determines the distribution of species across the planet, comprehensive data on species occurrence and environmental conditions are required. While some progress has been made in understanding the distribution of biodiversity at coarse spatial resolution, our knowledge of biodiversity at high spatial resolution is deficient [2]. In this study we consider biodiversity as the richness and turnover of taxonomic units, whether species or operational taxonomic units (OTUs) derived by meta-barcoding. Further, models of biodiversity for less well-known, but mega-diverse groups such as fungi and insects are almost non-existent across spatial scales [1]. Compared to targeted and systematic monitoring based on well-defined *a priori* hypotheses, surveillance data are often biased [e.g. temporal, spatial, taxonomic bias, 3] and therefore less appropriate and efficient for conservation management [4].

Recent developments in molecular techniques – in particular the extraction and sequencing of environmental DNA (eDNA) – hold the promise of more time-efficient sampling and identification of species [5, 6]. Further, eDNA enables the exploration of communities and organisms not easily recorded by traditional biodiversity assessment, such as soil-dwelling nematodes [7] . In fact, PCR-based methods combined with DNA sequencing have already provided valuable insight in the taxonomic diversity within complex environmental samples, such as soil [8-10] and water [e.g. 11, 12]. Due to the ongoing rapid development in DNA sequencing technology, with the emergence of next generation sequencing (NGS) techniques - generating billions of DNA sequences [13] - an environmental sample could now potentially be analyzed to a depth, which gives an almost exhaustive picture of the species composition at the site of collection. However promising, assessment of entire organismal communities from eDNA

samples is still in its early stages [6, 12, 14]. To assess the suitability and potential of eDNA data in complementing - or even replacing - traditional field survey data, tests on comprehensive data sets need to be done.

In sampling design, the ecospace framework [15] was followed. Thus, we aimed at a systematic sampling of the major aspects of environmental conditions (position), biotic resources (expansion) and spatio-temporal extent of biotopes (continuity). Environmental conditions and local processes may be considered a template shaping local biodiversity (e.g. through environmental filtering) [16, 17]. This aspect is reflected in ecospace position of sampled biotopes in abiotic environmental space. In addition to the physico-chemical conditions shaping abiotic gradients, particularly important to autotrophic organisms, we considered the presence and abundance of specific carbon resources, crucial to heterotrophic organisms, such as specialist herbivores, detritivores and saproxylic species [18]. The recording of organic carbon resources and structures, e.g. dead wood, dung and carcasses, is not often included in community studies, although the limited knowledge in the area [15, 19] speaks for further studies. Spatial and temporal processes at regional extent, such as extinction, speciation and migration, shape species pools and thereby set the limits to local richness and species composition [16, 17, 20]. In order to improve our understanding of biodiversity patterns, local and regional factors should be considered concurrently [17, 21]. Thus, in our study, we recorded spatial and temporal continuity of the local biotopes.

The aim of this study was to describe a comprehensive biodiversity monitoring protocol and evaluate its efficacy in describing environmental gradient variation and biodiversity across a large region. Here, we present the protocols we used in the project called *Biowide* to systematically and comprehensively map regional biodiversity and environmental heterogeneity

99 across Denmark - an area of limited geographical extent and with a relatively homogeneous
100 climate. We aimed to cover all of the major environmental gradients, including natural variation
101 in moisture, soil fertility and succession, as well as habitats under cultivation. Within this
102 environmental space, we performed a systematic and comprehensive sampling of the
103 environment and biodiversity. We combined traditional species observation and identification
104 with modern methods of biodiversity mapping in the form of sequencing of eDNA soil samples.

105 # Methods

106 ## Study area and site selection

107 We aimed to characterize biodiversity across the country of Denmark (Fig. 1a); a lowland area of
108 42,500 km^2 and an elevational range below 200 m. While there are some limestone and chalk
109 outcrops, there is no exposed bedrock in the investigated area. Soil texture ranges from coarse
110 sands to heavy clay and organic soils of various origins [22]. Land-use is dominated by arable
111 cultivation (61 %), most of it in annual rotation, while forest, most of which are plantations
112 established in the 19^{th} and 20^{th} centuries, and scrub cover approx. 17 %, natural and semi-natural
113 terrestrial habitats some 10 %, and freshwater lakes and streams 2 %. The remaining 10 % is
114 made up of urban areas and infrastructure [23, 24].

115 When selecting sites, we considered major environmental gradients, the area we would use as
116 a sampling unit, as well as practicalities of sampling. The selected observational unit was 40 × 40
117 m, which was a compromise between homogeneity and representativeness. We stratified site
118 selection according to the identified major environmental gradients, including the intensity of
119 human land use. We allocated 30 sites to cultivated habitats and 100 sites to natural and semi-
120 natural habitats. The cultivated subset represented major land-use categories and the natural
121 subset was stratified across natural gradients in soil fertility, soil moisture, and successional stage
122 from sparsely vegetated to closed canopy forest, (Appendix A). We deliberately excluded linear
123 features, such as hedgerows and road verges, urban areas with predominantly exotic plants as
124 well as saline and aquatic habitats, but included temporarily inundated heath and dune
125 depressions as well as wet mires.

126 The final set of 24 sampling strata consisted of six cultivated habitat types; three types of
127 fields (rotational, grass leys, set aside) and three types of plantations (beech, oak, spruce). The

remaining 18 natural strata constituted all factorial combinations of natural soil fertility (fertile and infertile), moisture (dry, moist and wet), and successional stage (low vegetation with bare soil, closed herb/scrub and forest) (Appendix A). These 24 strata were replicated in each of five geographical regions within Denmark (Fig. 1a). Finally, we included a subset of 10 sites placed within perceived hotspots for biodiversity in Denmark, selected subjectively by public polling among active natural history volunteers in the Danish nature conservation and nature management societies, but restricted so that each region held two hotspots. The result was 130 sites within 18 natural and 6 cultivated strata evenly distributed over the five geographic regions of Denmark (Table 1).

For the 18 natural habitat strata, site selection through stratified random sampling was guided by a large nation-wide data set of vegetation plots (n = 96,400 quadrats of 78.5 m^2, www.naturdata.dk) from a national monitoring and mapping project [25] and in accordance with the EU Habitats Directive [26]. We used plant indicator values to identify environmental conditions to select potential site candidates for the targeted strata. First, we calculated plot mean values for Ellenberg indicator values based on vascular plants species lists [27] and Grime CSR-strategy allocations of recorded plants [28], the latter recoded into numeric values following Ejrnæs & Bruun [29]. We initially excluded saline and artificially fertilized habitats by excluding plots with Ellenberg S > 1 or Ellenberg N > 6. We then defined stratification categories as: fertile (Ellenberg N 3.5-6.0), infertile (Ellenberg N < 3.5), dry (Ellenberg F < 5.5), moist (Ellenberg F 5.5-7.0), wet (Ellenberg F > 7.0), early succession (Grime R > 4 and Ellenberg L > 7 or > 10 % of annual plants), late succession (mapped as forest), mid succession (remaining sites).

149 To reduce transport time and costs, all 26 sites within each region were grouped into three
150 geographic clusters (Fig. 1a). The nested sampling design was also considered an opportunity to
151 take spatially structured species distributions into account [30].
152 The procedure for site selection involved the following steps:
153 1) Designation of three geographic clusters within each region with the aim to cover all
154 natural strata while a) keeping the cluster area below 200 km^2 and b) ensuring high
155 between-cluster dispersion in order to represent the geographic range of the region. In
156 practice, hotspots were chosen first, then clusters were placed with reference to the
157 highest ranking hotspots and in areas with a wide range of strata represented in our
158 national monitoring plot data [NOVANA, 31].
159 2) Representing 24 strata in each region by selecting 8-9 potential sites in each cluster.
160 Natural strata were selected from classified field-plot data whereas cultivated strata were
161 assumed omnipresent and used as buffers in the process of completing the non-trivial task
162 of finding all strata within three restricted cluster areas in each region.
163 3) Negotiating with land owners and, in case of disagreement, replacing the preferred site
164 with an alternative site from the same stratum.
165 After each of the 130 sites were selected using available data we established each 40 × 40 m
166 field site in a subjectively selected homogenous area that accounted for topography and
167 vegetation structure. Each site was divided into four 20 × 20 m quadrants and from the center of
168 each quadrant a 5 m radius circle (called a plot) was used as a sub-unit for data collection to
169 supplement the data collected at site level (40 × 40 m) (Fig. 1b).

170 **Collection of biodiversity data**

For each of the 130 field inventory sites, we aimed to make an unbiased and representative assessment of the multi-taxon species richness. Data on vascular plants, bryophytes, lichens, macrofungi, arthropods and gastropods were collected using standard field inventory methods (Appendix B). For vascular plants, bryophytes and gastropods, we collected exhaustive species lists. For the remaining taxonomic groups that are more demanding to find, catch, and identify, we aimed to collect a reproducible and un-biased sample through a standardized level of effort (typically one hour). Each site was carefully examined for lichens and macrofungi assessing various substrates (soil, herbaceous debris, wood, stone surfaces and bark of trees up to 2 m). For fungi, we visited each site twice during the main fruiting season in 2014 – in August and early November – and once during the main fruiting season in 2015 – from late August to early October. Specimens that were not possible to identify with certainty in the field were sampled and, when possible, identified in the laboratory. For arthropod sampling, a standard set of pitfall traps (including meat-baited and dung-baited traps), yellow Möricke pan traps and Malaise traps were operated during a fixed period of the year. In addition, we used active search and collection methods, including sweep netting and beating as well as expert searches for plant gallers, miners and gastropods. Finally, we heat-extracted collembolas and oribatid mites from soil cores. Due to the limited size of the sites relative to the mobility of mammals, birds, reptiles and amphibians, data on these groups were not recorded. Taxonomic data will be transferred to the Global Biodiversity Information Facility (GBIF, http://www.gbif.org/) and specimens to the Natural History Museums, when the project ends in 2017. For further details on methods for collection of biodiversity data see Appendix B.

Collection of eDNA data

We used soil samples collected from all 130 sites for the eDNA inventory. At each site, we sampled soil cores in grids embedded in the 9 × 9 plots (81 soil cores per site) and pooled the collected samples after removal of coarse litter. We homogenized the soil by mixing with a drilling machine mounted with a mixing paddle. A subsample of soil was sampled from the homogenized sample and DNA was extracted for marker gene amplification and sequencing [14]. We chose the MiSeq platform by Illumina for DNA sequencing because, relative to other platforms (e.g., 454 b Roche), it produces 15 times the sequence output (approx. 15 000 000 reads). MiSeq is adapted to amplicon sequencing [32]. To our knowledge such comprehensive regional inventories of soil communities has not been carried out before.

For further details on methods for eDNA data and considerations on eDNA species richness and community composition measures see Appendix B.

Site environmental data

We have followed the suggestion in Brunbjerg et al. [15] to describe the fundamental requirements for biodiversity in terms of the ecospace (position, expansion and spatio-temporal continuity of the biotope).

Position

To assess the environmental variation across the 130 sites, we measured a core set of site factors that described the abiotic conditions at each site. Environmental recordings and estimates included soil pH, total soil carbon (C, g/m^2), total soil nitrogen (N, g/m^2) and total soil phosphorus (P, g/m^2), soil moisture (% volumetric water content), leaf CNP (%), soil surface temperature (°C) and humidity (vapour pressure deficit), air temperature (°C) and light intensity (Lux). For further details on methods for collection of the abiotic data see Appendix B.

215 ***Expansion***

216 We collected measurements that represented the organic C resources species consume as well as
217 organic C structures that species can use as habitat. While many invertebrates are associated with
218 other animals, in order to accomplish accurate sampling of the focal species we restricted our
219 mapping of carbon space to the variation in live and dead plant tissue, including dung. We
220 measured litter mass (g/m^2), plant species richness, vegetation height (of herb layer, cm), cover
221 of bare soil (%), bryophyte cover (%) and lichen cover (%), dead wood volume (m^3/site),
222 dominant herbs, the abundance of woody species, the number of woody plant individuals, flower
223 density (basic distance abundance estimate, [33]), density of dung (basic distance abundance
224 estimate), number of carcasses, fine woody debris density (basic distance abundance estimate),
225 ant nest density (basic distance abundance estimate), boulders density and water puddle density
226 (basic distance abundance estimate). For further details on methods for collection of expansion
227 data see Appendix B.

228 ***Mapping of temporal and spatial continuity***

229 For each site, we inspected a temporal sequence of aerial photos (from 1945 to 2014) and
230 historical maps (1842-1945) starting with the most recent photo taken. We defined temporal
231 continuity as the number of years since the most recent major and documented land use change.
232 The year in which a change was identified was accepted as time for 'break in continuity'. To
233 estimate spatial continuity, we used ArcGIS to construct four buffers for each site (500 m, 1000
234 m, 2000 m, 5000 m). Within each buffer we estimated the amount of habitat similar to the site
235 focal habitat by visual inspection of aerial photos with overlays representing nation-wide
236 mapping of semi-natural habitat. For further details on methods for collection of continuity data
237 see Appendix B.

Analyses

To illustrate the coverage of the three main gradients (moisture, fertility, and successional stage) spanned by the 130 sites, Ellenberg mean site values (mean of mean Ellenberg values for the four 5m radius quadrats within each site) for soil moisture (Ellenberg F), soil nutrients (Ellenberg N) and light conditions (Ellenberg L) were plotted relative to Ellenberg F, N and L values for a reference data set of 5m radius vegetation quadrats (47,202 from agricultural, semi-natural and natural open vegetation and 12,014 from forests (www.naturdata.dk) [25]. Mean Ellenberg values were only calculated for quadrats with more than five species and 95 percentile convex hull polygons where drawn for the reference data set as well as the Biowide data set.

We assessed the regional coverage of species in the project, with reference to the number of known species from Denmark according to the taxonomic database Allearter (www.allearter.dk). This portal represents the most up-to-date list of species known from Denmark. Coverage (or sample completeness) was estimated for each taxonomic group across sites (Biowide coverage) as well as for each site individually (site coverage) for species groups with abundance data (Diptera, Coleoptera and Araneae) by comparing the number of species found to the estimated species richness of the sample using the iNEXT R-package [34].

To further evaluate how well we covered the environmental gradient for our inventory, we related community composition to the measured environmental variables (abiotic and biotic) based on a Nonmetric Multidimensional Scaling (NMDS) analyses in R v. 3.2.3 [35] using the vegan R-package [36] and the plant species × site matrix as well as the macrofungi species × site matrix. Abiotic and biotic variables were correlated with ordination axes to facilitate interpretation.

260 To illustrate and substantiate the adequacy of the eDNA sampling protocol and subsequent
261 laboratory protocols, we correlated basic biodiversity measures of community composition
262 (NMDS axes) and richness for plant eDNA (ITS2 marker region) with the same measures for our
263 observed plant data (see Appendix B for detailed methods).
264

Results

The 130 sites were distributed in 15 clusters nested within five regions across Denmark (Fig. 1a). The measured variables differed according to the initial stratification of sites based on simple indicators (Table 1, Fig. 2a, b, ranges of measured variables in Appendix C). Managed sites (plantations and agricultural fields) revealed little variation in soil moisture (Fig. 2b). The Hotspots spanned the full variation of natural sites regarding fertility, moisture and successional stage (Fig 2b).

The selected 130 sites covered the main gradients reflected by a huge reference dataset from a national monitoring program (Fig. 3) as judged from a vegetation-based calibration of site conditions regarding moisture, fertility and succession (light intensity). Biowide data seemed to increase the upper range of the fertility gradient, which can be explained by the inclusion in Biowide of rotational fields that were not included in reference data (Fig. 2b, 3).

The environmental expansion of ecospace, which was measured as the amount and differentiation of organic carbon sources, varied among habitat types with high litter mass in plantations and late successional habitats, high plant species richness in early and mid-successional habitats, high dung density in open habitats (early successional and fields) and high amounts of dead wood in late successional habitats (Fig. 4). Spatial and temporal continuity varied for the 130 sites with less spatial continuity at larger buffer sizes (Fig. 5). The number of species found per site differed with taxonomic group with the highest number for macrofungi and lowest for bryophytes and lichens (Fig. 6).

We collected 2040 species of macrofungi (corresponding to 62 % of the number of macrofungi recorded in Denmark), 202 lichens (20 %), 663 vascular plants (42 %) and 254 bryophytes (41 %) during the monitoring period. We collected 75 species of gastropods (75 %),

334 spiders (59 %), 105 hoverflies (36 %), 126 carabid beetles (38 %) and 203 galler and miner species (21 %). For all groups except macrofungi, the number of species found was highest in natural sites (90 sites of 130), but across taxonomic groups, plantations and agricultural fields harbored unique species – plantations were particularly important in harboring unique species of macrofungi (Table 2). The taxonomic sample coverage calculated by rarefaction within the 130 sites was high overall (range: 0.86-0.99), but highest for gastropods and spiders and lowest for gallers and miners (Table 2).

The inventory was unprecedented in detail and resulted in a total of 118 new macrofungi, 1 new lichen and 32 new invertebrate species (of which 12 were gallers and miners and 3 spiders) that had not previously been documented in Denmark (Table 2).

The NMDS ordination (3-dimensional, final stress = 0.102) accounted for 81 % of the variation in plant species composition and 72 % of the variation in macrofungal species composition (3-dimensional, final stress = 0.146). The major gradients in plant species composition of the 130 sites correlated strongly with soil fertility (NMDS axis 1 strong correlation with soil N, P and pH), successional stage (NMDS axis 2 strong correlation with light intensity and opposite correlation with litter mass and number of large trees) and soil moisture (NMDS axis 3 strong correlation with measured soil moisture), reflecting the gradients that the sites were selected to cover (Fig. 7, see correlation matrix for the rest of the environmental variables in Appendix D). Macrofungal species composition showed the same gradients, however succession and fertility swapped with succession as primary gradient (NMDS1) and fertility as secondary gradient (NMDS2). NMDS axis 3 reproduced a strong correlation with soil moisture.

310 Spearman Rho correlations between observational plant species richness and eDNA OTU
311 'richness' as well as observational plant community composition (as represented by NMDS axes
312 1-3) and eDNA OTU composition were both strong and confirmative for a recovery of plant
313 diversity by soil-derived DNA ($R^2_{richness}$=0.652, $R^2_{composition}$=0.577-697, Fig. 8).

Discussion

Using ecospace as conceptual framework [15], we developed a protocol for mapping terrestrial biodiversity at a regional level and covering numerous, mega-diverse taxa. Across the 130 surveyed sites, covering a tiny fraction (0.0005 %) of the total area of Denmark, we observed a total of ~5 700 species, of which 150 represented new species records for the country, and 20-75 % of known regional species number of species depending on taxonomic group. Our data indicated that the sampling at 130 sites sufficiently covered the known local and regional environmental variation of Denmark and also delivered a good coverage of biodiversity at the spatial scale of sites – even for diverse groups of invertebrates and fungi. Finally, the study demonstrates that eDNA data, once properly curated (Frøslev et al. submitted), may be used as an important supplement to classical biodiversity monitoring.

Environmental filtering is an important process in community assembly, reflecting the prominent role of niche-differentiation in evolution [37]. The most obvious design principle for a biodiversity inventory is, therefore, to stratify sampling according to major abiotic and biotic environmental gradients [e.g. 38]. We found a close correspondence between the variation in average Ellenberg values at our sites and those extracted from a very large vegetation database comprising vascular plant species lists from a national monitoring program. This indicates an almost complete gradient coverage in our study and allows us to generalize relationships between environment and biodiversity derived from local measurements across gradients to a large spatial extent. Although the use of stratified random sampling implies a biased representation of rare and common environmental conditions, complete random sampling would have led to limited representation of natural biotopes and their disproportionate contribution to the total biodiversity may have been missed.

337 While the ecospace framework helped structure our sampling, it also proved challenging with
338 respect to decisions about site area and homogeneity (related to ecospace position), recording of
339 carbon resources (assessing ecospace expansion) and definition of temporal and spatial
340 continuity. Ideally, abiotic and biotic conditions should be homogenous across a site in order to
341 ensure that site measurements reflect the abiotic position and biotic expansion [15]. Thus, site
342 area was a trade-off between homogeneity (small area) and representativeness (large area) and
343 across long environmental gradients, homogeneity and representativeness may vary among for
344 example, grassland, heathland, and forest. Similarly, while counting the number of different
345 plant species is easy, accounting for the relative contribution of each species to total biomass and
346 measuring the availability of different pools of wood, woody debris, litter, dung, flowers and
347 seeds is much harder. Finally, spatial and temporal continuity is hard to quantify due to data
348 limitations and because past soil tillage, fertilization, or other land management or disturbance
349 regimes have not been recorded and must be inferred. In addition, an unambiguous definition of
350 continuity breaks is impossible given that most land use changes and derived community
351 turnover occur gradually over time. We estimated spatial continuity using broad habitat classes at
352 a range of scales (500 m, 1000 m, 2000 m, 5000 m) acknowledging that the dependency on
353 spatio-temporal continuity depend on the mobility, life history and habitat specificity of different
354 species. Our estimates of temporal continuity were also limited by the availability of aerial
355 photographs and maps, which while not perfect, is good relative to other parts of the world.
356 Despite these constraints, our estimates of spatial and temporal continuity varied among sites and
357 were uncorrelated, which allowed us to statistically test for their relative roles.
358 We aimed for equal sampling effort per site in terms of trapping and searching time.
359 However, this was challenged by an array of practicalities. The preferred species sampling

methods varied among taxonomic groups [39, 40] and despite our application of a suite of methods, including passive sampling in pitfall traps and Malaise traps, baited traps, soil core sampling and active search, our taxonomic coverage is still incomplete (e.g. aphids, Phorid flies and other species-rich groups living in the canopy are inevitably under-sampled). Our budget also forced us to be selective with the identification of the most difficult species groups, in particular within Hymenoptera and Diptera. Among identified groups, species coverage ranged between 20 and 75 % of all species known to Denmark, which is quite satisfactory. Invertebrate sampling and identification is extremely time consuming and relies on rare taxonomic expertise. We spent more than half of the inventory budget on invertebrate sampling and identification, and yet the site coverage remained modest in some sites across all invertebrate groups. Invertebrates constitute by far the largest fraction of the total biota and, for many species, the adult life stage is short-lived, highly mobile, and the range of active species varies with season [41, 42]. Trapping also implies a certain risk of suboptimal placement or vandalism by visiting humans, domestic livestock or wild scavengers. The resulting number of invertebrate species per site is relatively high and revealed a considerable variation, which gives ample opportunity for comparative analyses. Furthermore, the high coverage of invertebrates across the full 130 sites indicated that the variation in site conditions and biota was adequately sampled. The high number of new species for Denmark, particularly macrofungi, can most likely be attributed to the effort, but also to the inclusion of habitat types that would otherwise have been avoided or overlooked during voluntary monitoring. This underpins the qualitative differences between surveillance data and targeted monitoring [3].

Although methods for DNA extraction, amplification, sequencing and bioinformatic processing are continuously improved and may lead to better biodiversity metrics from

environmental samples, collecting representative samples from larger areas with unevenly distributed species remains a challenge. We pooled and homogenized large amounts of soil, followed by extraction of intracellular as well as extracellular DNA, from a large subsample, to maximize diversity coverage within a manageable manual workload. Biodiversity metrics based on plant DNA were correlated to the same metrics for observational plant data. This indicates that the procedure for sampling, DNA extraction and amplification can be assumed to be adequate for achieving amplicon data to quantify variation in biodiversity across wide ecological and environmental gradients for plants, but most likely also for other organisms present in the soil. These methods are promising for biodiversity studies of many organism groups that are otherwise difficult to sample and identify (e.g. nematodes, fungi, protists, and arthropods). High throughput sequencing methods produce numerous errors [e.g., 43, 44] and it has been suggested that richness measures should be avoided altogether for HTS studies [45]. Despite the remaining challenge of relating genetic units to well-known taxonomic entities, our results along with those presented in Frøslev et al. (in review) indicate that reliable metrics of α-diversity and community composition are achievable. With respect to taxonomic annotation, reference databases are far from complete and the taxonomic annotation of reference sequences are often erroneous. Furthermore, for many groups of organisms, we have still only described and named a fraction of the actual species diversity, and the underlying genetic diversity within and between species is largely unknown for most taxa, leading to uncertainties in OTU (/species) delimitation and taxonomic assignment of sequence data. This also means that ecological interpretation of OTU/species assemblages assessed by eDNA is largely impossible as there is little ecological knowledge that can be linked to OTUs. Thus, for eDNA based biodiversity assessment to further mature, molecular biologists, ecologists and taxonomists need to work closely together to

406 produce well-annotated reference databases and to relate unnamed OTUs to well-described
407 ecospaces. Our environmental samples for eDNA, including soil and litter samples as well as
408 extracted DNA will be preserved for the future. This material represents a unique resource for
409 the further development of methods within ecology and eDNA. As more efficient technologies
410 become available in the future, it will be possible to process this material at an affordable cost
411 and derive further insights on the relationship between traditional species occurrence, OTU data
412 and environmental variation.

413 Conclusion

414 We have presented a generic protocol to obtain a representative, un-biased sample of multi-taxon
415 biodiversity stratified with respect to the major abiotic gradients. By testing and evaluating the
416 protocol, we conclude that it is operational and that observed biodiversity variation may be
417 accounted for by the measured abiotic and biotic variables. We believe that the ecospace
418 concept, on which this protocol is developed, can be successfully up-scaled and applied to
419 biodiversity studies at regional to continental scale. Despite the obvious advantages of eDNA
420 data (economically and logistically), barcode reference libraries are as yet far too incomplete.
421 Thus, combining classical taxonomic identification with eDNA sampling proves a promising
422 approach for biodiversity science.

Additional files

Additional file 1: Appendix A: Site characteristics for each of the 130 40 ×40 m sites.

Additional file 2: Appendix B: Protocols for data collection**.**

Additional file 3: Appendix C: Ranges of environmental (abiotic and biotic) variables measured within the 130 sites as well as species richness of various taxonomic groups.

Additional file 4: Appendix D: Correlation matrix for NMDS axes 1, 2 and 3 and environmental variables

Authors' contributions

AKB, HHB, AC, TGF, TL, MDDH and RE conceived and designed the study. AKB, HHB, LB, KF, TGF, IG, TL, GN, LS, US and RE conducted field work. AKB, RE and TGF analyzed the data and prepared the figures. LB, KF, IG, MDDH, TL, LS, US and HHB sorted and identified specimens. AKB, HHB, LB, ATC, KF, IG, MDDH, TTH, TGF, TL, GSN, LS, US and RE wrote the manuscript. All authors have read and approved the final version of the manuscript.

Acknowledgements

We thank Ako O. Mirza for plant and soil lab work, Vagn Alstrup and Roar Skovlund Poulsen for lichen monitoring, volunteers that have helped in data collection, land owners, Karl-Henrik Larsson for aid in identifying critical corticioid fungi, Leif Örstadius for identifying Psathyrella collections. In regard to identification of invertebrate specimens we would like to thank Henning Petersen for identifying Springtails (Collembola), Hjalte Kjærby for identifying Grasshoppers (Orthoptera) and Harvestmen (Opilliones), Kåre Fog for identifying snails (Gastropoda), Kåre Würtz Sørensen for identifying various Wasps (Symphyta, Spechidae, Crabronidae, and Vespidae), Lars Dyhrberg Bruun for identifying spiders (Aranea), Lars Brøndum for identifying Hoverflies (Syrphidae), Carrion Beetles (Silphidae) as well as Scarabs (Scarabidae), Lars Skipper for identifying True bugs (Heteroptera), Maja Møholt for identifying Dung beetles (*Aphodius, Onthophagus* and Geotrupidae) and Cantharidae, Marianne Graversen for identifying Longhorn beetles (Cerambycidae) and Ladybugs (Coccinellidae), Peter Wiberg-Larsen for identifying Caddisflies (Trichoptera), Mathias Holm for identifying True weevils and Seed weevils (Curculionidae and Apionidae), Monica Aimeé Harlund Oyre for identifying various Dipterans (Syrphida, Tachinadae, Stratiomyidae, Acroceridae, Rhagionidae, Tephritidae, Plastytomatidae, Asilidae) as well as Strepsipterans (Strepsiptera) and Book- and Barklice

(Psocoptera), Morten D. D. Hansen for identifying Bees (Apoidea), Carrion beetles (Silphidae), Click beetles (Elateridae), Scarabs (Scarabaeidea) and Dung beetles (*Aphodius*, *Onthophagus* and Geotrupidae), Ole Fogh Nielsen for identifying net-winged insects (Neuroptera) and Strepsipterans (Strepsiptera), Oskar Liset Pryds Hansen and Emil Skovgaard Brandtoft for identifying Ground beetles (Carabidae), Sofie Amund Kjeldgaard and Steffen Kjeldgaard for identifying Owlet moths (Noctuidae), Mathias G. Skytte for identifying Rove beetles (Staphyllinidae), Simon Haarder for identifying galling and mining arthropods, and Ulrik Hasle Nielsen for identifying Cicadas (Cicadoidea) as well as numerous other volunteers. In regard to carrying out the eDNA lab work we would like to thank Anne Aagaard Lauridsen, Sarah Mak, Stine Raith Richter, Carlotta Pietroni and Ida Broman Nielsen**.**

Tables

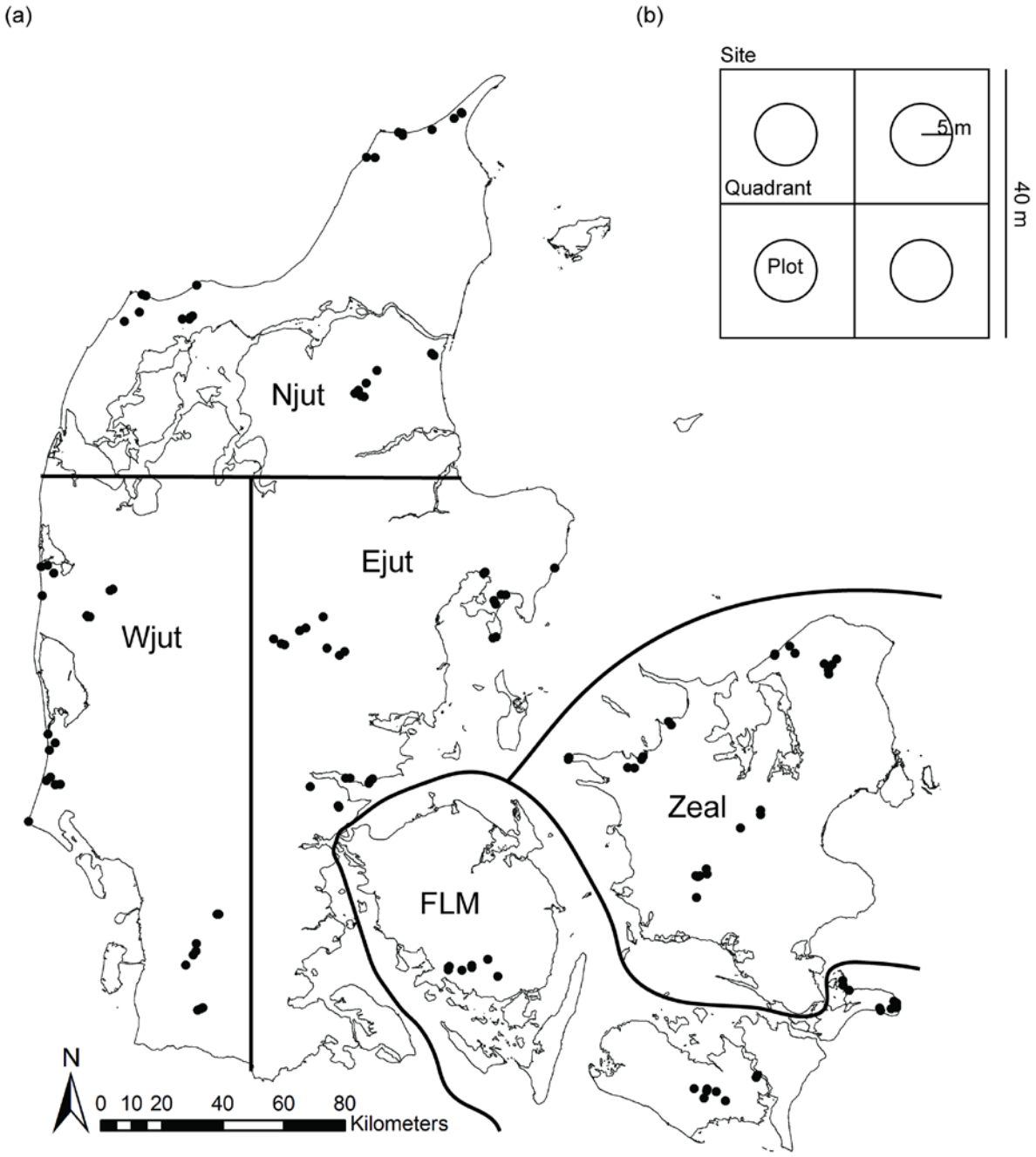
Table 1: Stratification of sites in survey. Number of sites (N) within each habitat category (fields, plantations, hotspots and natural gradients) and type.

Habitat	Category	Type	N
Cultivated	Fields	Rotational	5
		Grass ley	5
		Set aside	5
	Plantations	Beech	5
		Oak	5
		Spruce	5
Natural/uncultivated	Hotspots	NA	10
	Natural	Fertile/Dry/Early	5
	Natural	Fertile/Dry/Mid	5
	Natural	Fertile/Dry/Late	5
	Natural	Fertile/Moist/Early	5
	Natural	Fertile/Moist/Mid	5
	Natural	Fertile/Moist/Late	5
	Natural	Fertile/Wet/Early	5
	Natural	Fertile/Wet/Mid	5
	Natural	Fertile/Wet/Late	5
	Natural	Infertile/Dry/Early	5
	Natural	Infertile/Dry/Mid	5
	Natural	Infertile/Dry/Late	5
	Natural	Infertile/Moist/Early	5
	Natural	Infertile/Moist/Mid	5
	Natural	Infertile/Moist/Late	5
	Natural	Infertile/Wet/Early	5
	Natural	Infertile/Wet/Mid	5
	Natural	Infertile/Wet/Late	5

Table 2: Number of species per taxonomic group found during the project. Number of species per taxonomic group found in natural sites (n=90), hotspots (n=10), plantations (n=15) and arable land (n=15). Number of unique species for each stratum and taxonomic group is given in parentheses. Regional coverage (number of species found divided by the number of known species for Denmark for each taxonomic group, www.allearter.dk), Biowide coverage (iNEXT function) and site coverage (site coverage range for 130 sites for invertebrates with abundance data (iNEXT function)) are given (www.allearter.dk) as well as the number of new species for Denmark found during the project.

	Habitat type					Coverage			
	Total	Natural	Hotspots	Arable	Plantations	Regional	Biowide	Site	New species (Denmark)
Sites (N)	130	90	10	15	15				
Plants	719	601 (225)	330 (21)	192 (47)	131 (2)	0.36 (719/2017)	0.972	P/A	0
Bryophytes	254	221 (106)	96 (11)	20 (3)	78 (4)	0.41 (254/621)	0.974	P/A	0
Macrofungi	2040	1550 (1013)	620 (134)	146 (19)	557 (131)	0.62 (2040/3274)	0.920	P/A	118
Lichens	202	183 (92)	76 (9)	19 (5)	58 (3)	0.20 (202/1035)	0.955	P/A	1
Spiders	334	308 (102)	145 (5)	125 (5)	126 (12)	0.59 (334/570)	0.983	0.514-0.989	3
Hoverflies	105	93 (43)	36 (3)	45 (6)	20 (2)	0.36 (105/296)	0.969	0.053-1.000	0
Gastropods	75	72 (18)	42 (0)	19 (1)	38 (2)	0.75 (75/100)	0.991	P/A	0
Carabids	126	107 (46)	34 (3)	51 (15)	35 (1)	0.38 (126/336)	0.960	0.039-0.977	0
Gallers/miners	203	169 (108)	48 (10)	19 (6)	41 (16)	0.21 (203/968)	0.863	P/A	12

476 # Figures



477
478 **Figure 1:** a) Map of Denmark showing the location of the 130 sites grouped into 15 clusters

479 within five regions (Njut: Northern Jutland, Wjut: Western Jutland, Ejut: Eastern Jutland, FLM:

480 Funen, Lolland, Møn, Zeal: Zealand). b) Site layout with four 20 × 20 m quadrants each

481 containing a 5 m radius circle (plot).

482

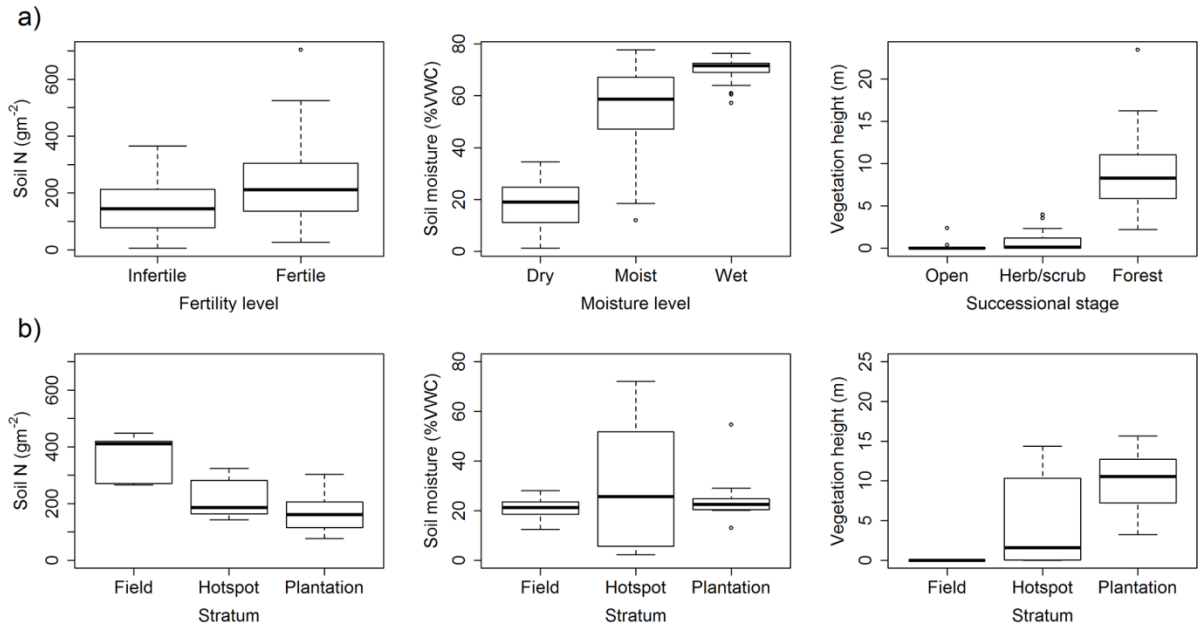


Figure 2: Validation of the stratification scheme used in site selection. Boxplots of measured values of nutrient levels (soil N g/m^2), moisture levels (trimmed site mean % Volumetric Water Content (VWC)), and vegetation height (mean LIDAR canopy height (m)) for the a) 90 natural sites of different strata (fertility level: infertile, fertile; moisture level: dry, moist, wet; successional stage: early (open), mid (herb/scrub), late (forest); and b) the 15 plantations, 15 fields and 10 hotspots.

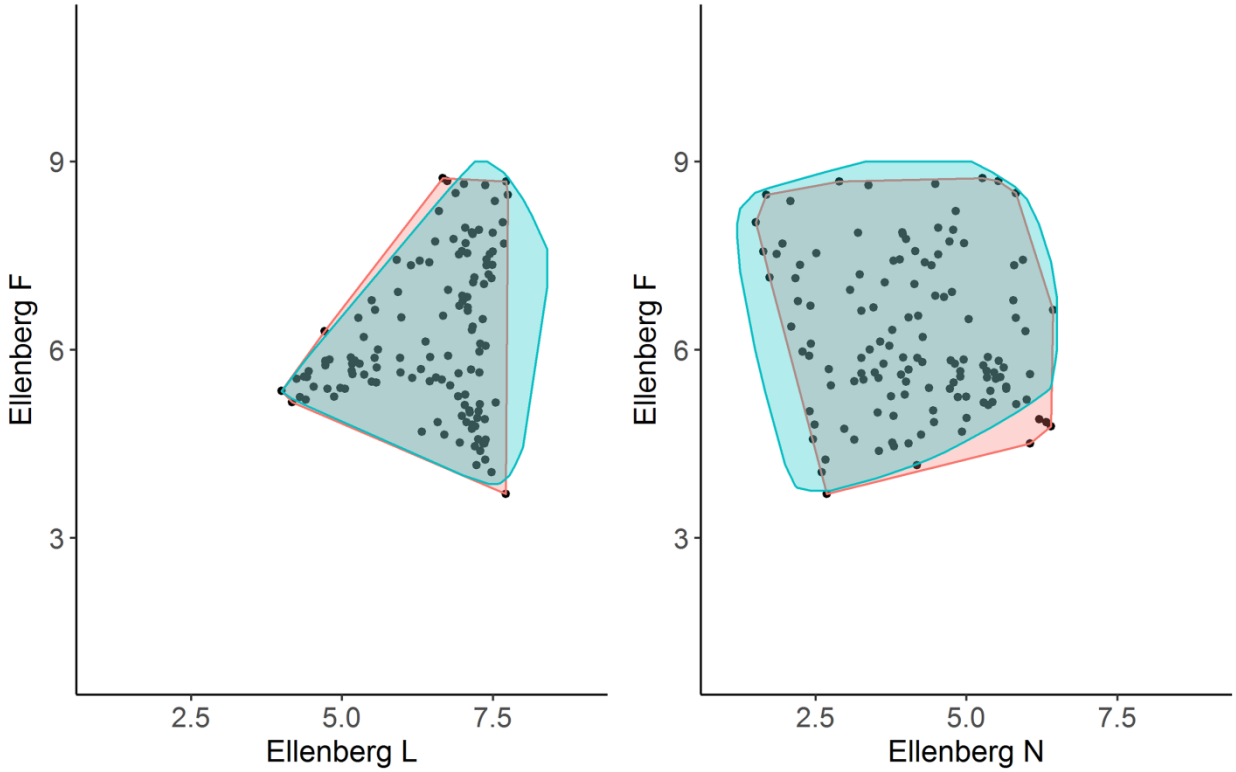


Figure 3: 95 percentile convex hull plots of Ellenberg F, L and N values from a reference data set (www.naturdata.dk) of open and forest habitat types (blue, n= 59 227) as well as the data set used in this study, Biowide (red, n=130). Black dots represent Ellenberg values of the 130 Biowide sites.

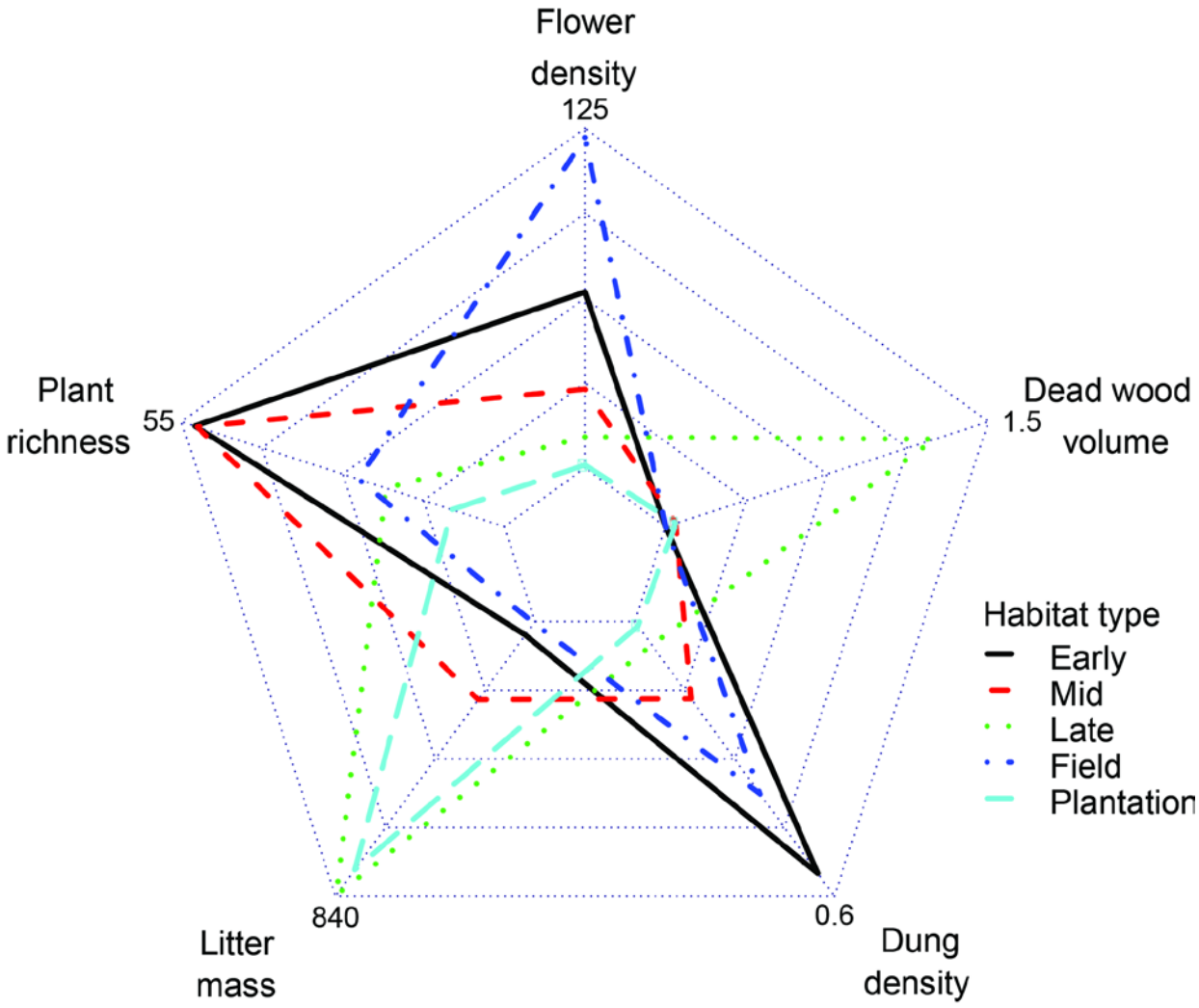


Figure 4: Habitat mean values for various carbon resources in the 130 40 × 40 m sites. Volume of dead wood (m^3/ha), density of dung (cow, sheep, deer, horse, hare) (number/m^2), summed flower density in April, June and August (number/m^2), litter mass (g/m^2) and plant species richness per site are depicted for natural habitat types (early, mid and late successional stage), fields and plantations.

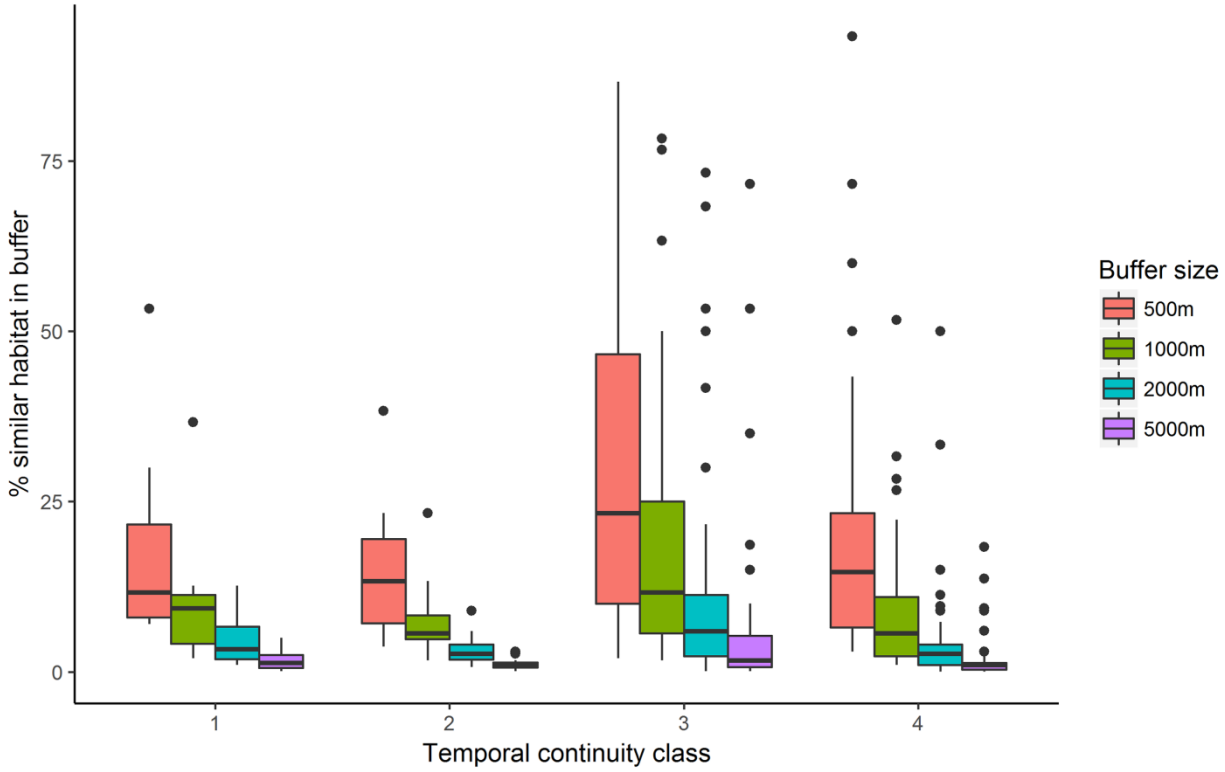
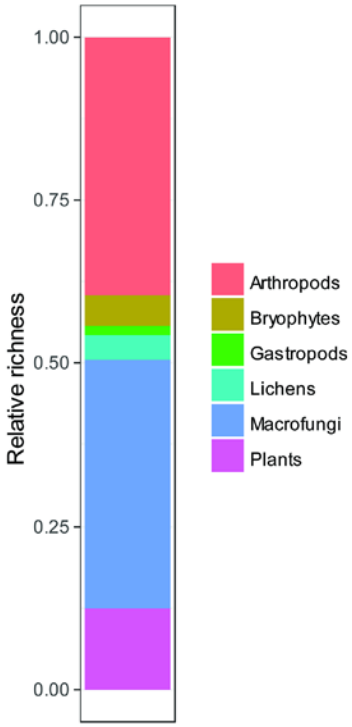


Figure 5: Temporal and spatial continuity for the 130 Biowide sites. Temporal continuity is represented on a 4-level ordinal scale: 1: < 15 years of continuity, 2: 15-44 years of continuity, 3: 45-135 years of continuity and 4: >135 years of continuity (or continous on the oldest available map). Spatial continuity represents the amount of similar habitat (%) as the focal site habitat within four different buffer sizes (500m, 1000m, 2000m, 5000m).



507

508 **Figure 6**: Relative richness of arthropods, bryophytes, gastropods, lichens, macrofungi and

509 vascular plants across all 130 sites.

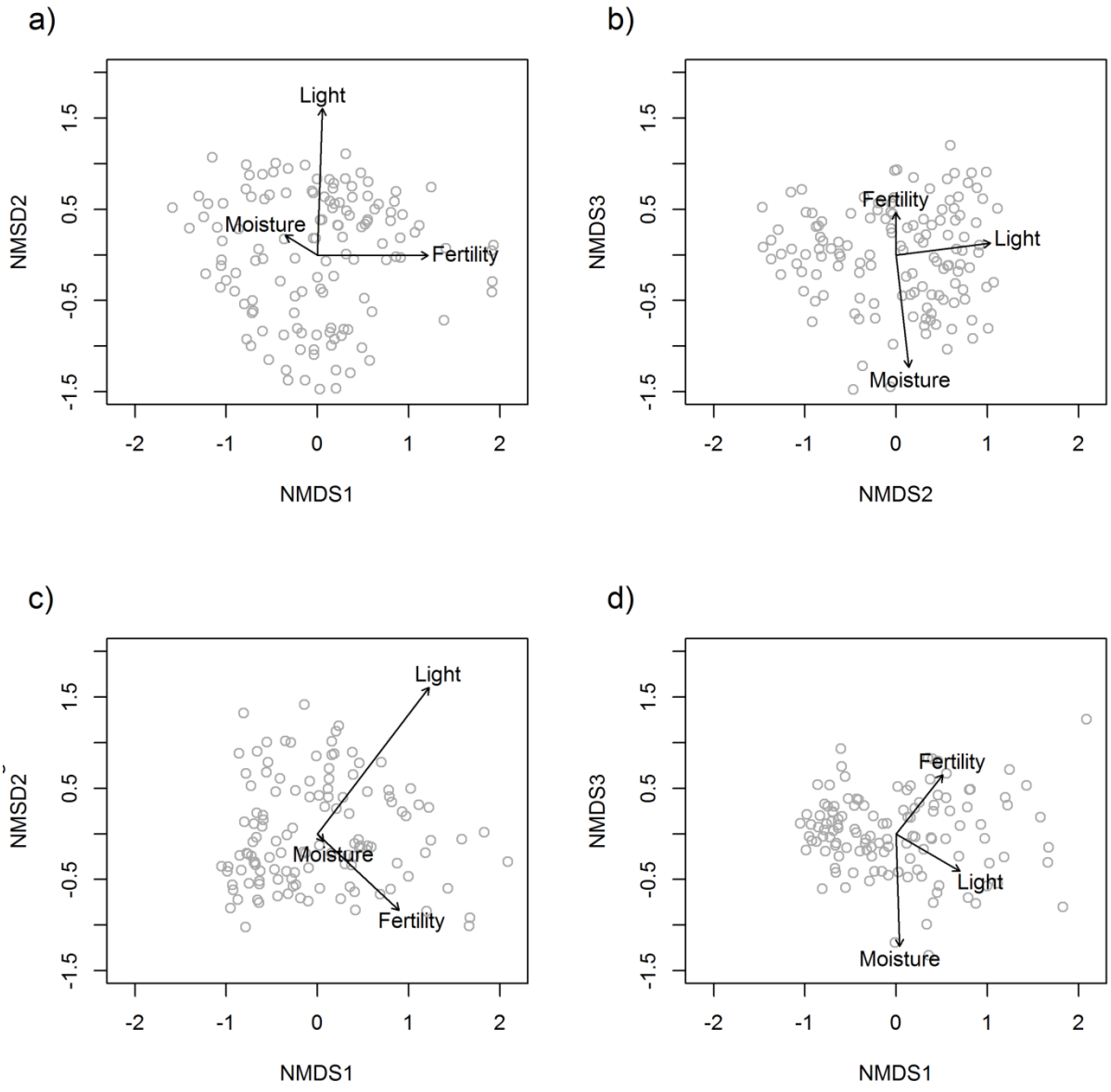


Figure 7: Three dimensional NMDS plots for plants with a) showing axis 2 against axis 1 and b) showing axis 3 against axis 2 and fungi with c) showing axis 2 against axis 1 and d) showing axis 3 against axis 1. The three main gradients used for selecting the 130 sites (fertility, moisture, successional stage) are overlaid as arrows (from an envfit analyses in the R package Vegan). The ordinations are based on plant species lists from the 130 sites a) & b) or macrofungi species lists from the 124 sites with more than five species c & d) and the arrows reflect soil moisture measured using a soil moisture meter, fertility measured as soil N and light measured as light

518 intensity using HOBO loggers. The ordination plots illustrate that the community composition of
519 vascular plants and macrofungi actually reflect the main gradients the sites were selected to
520 cover. The scatter of dots shows the variation in abiotic conditions across the 130 sites.
521 Correlations and p-values can be seen in Appendix D.
522

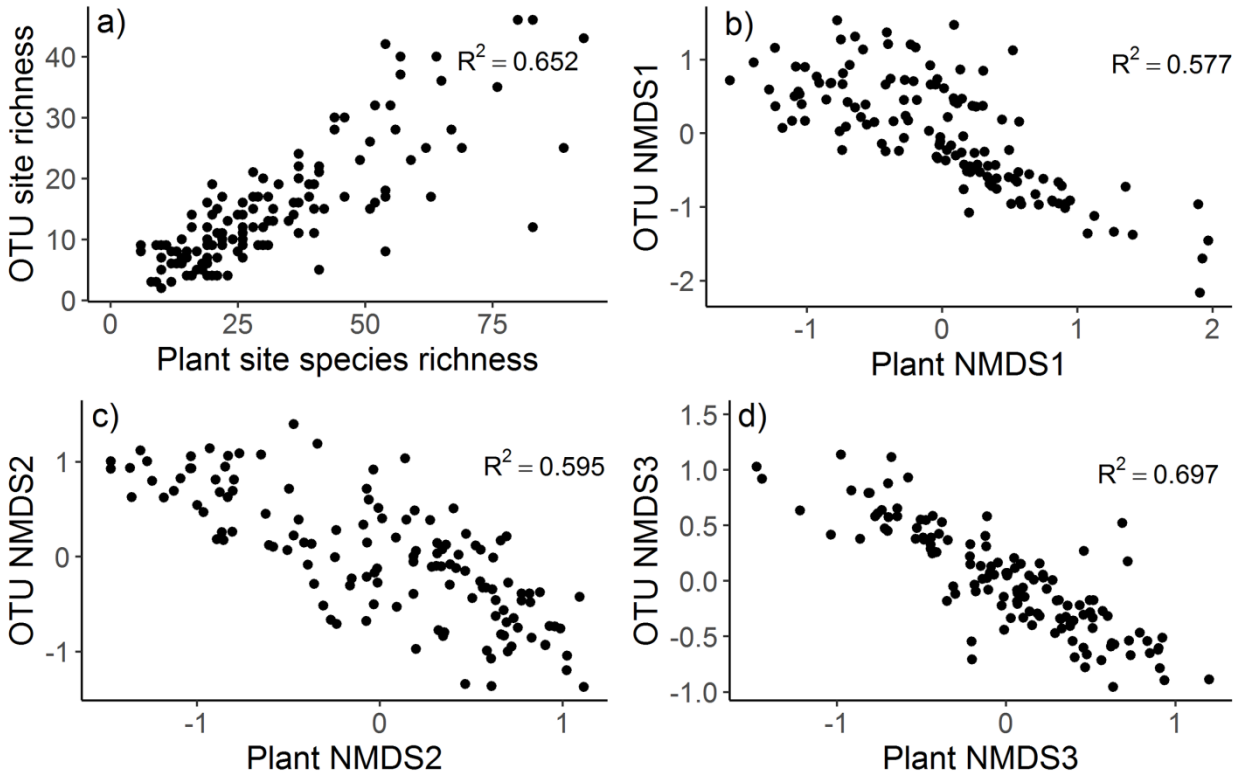


Figure 8: Correlation between a) observed site plant species richness and OTU site 'richness' for the 130 sites (Spearman Rho: R^2=0.652, S = 70457, p-value < 0.001), b-d) observed site plant community composition and OTU community composition for the 130 sites b) NMDS axes 1 (Spearman Rho: R^2=0.576, S = 644210, p-value < 0.001), c) NMDS axes 2 (Spearman Rho: R^2=0.594, S = 648480, p-value < 0.001), and d) NMDS axes 3 (Spearman Rho: R^2=0.697, S = 671850, p-value < 0.001).

References

1. Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. How many species are there on Earth and in the ocean? PLoS Biol 2011; 9:e1001127.
2. Jetz W, McPherson JM, Guralnick RP. Integrating biodiversity distribution knowledge: toward a global map of life. Trends Ecol Evol 2012; 27:151-159.
3. Isaac NJB, van Strien AJ, August TA, de Zeeuw MP, Roy DB. Statistics for citizen science: extracting signals of change from noisy ecological data. Methods in Ecology and Evolution 2014; 5:1052-1060.
4. Nichols JD, Williams BK. Monitoring for conservation. Trends Ecol Evol 2006; 21:668-673.
5. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol 2012; 21:2045-2050.
6. Thomsen PF, Willerslev E. Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. Biological Conservation 2015; 183:4-18.
7. Porazinska DL, Giblin-Davis RM, Esquivel A, Powers TO, Sung WAY, Thomas WK. Ecometagenetics confirm high tropical rainforest nematode diversity. Mol Ecol 2010; 19:5521-5530.
8. Yoccoz NG, Brathen KA, Gielly L, Haile J, Edwards ME, Goslar T, von Stedingk H, Brysting AK, Coissac E, Pompanon F *et al*. DNA from soil mirrors plant taxonomic and growth form diversity. Mol Ecol 2012; 21:3647-3655.
9. Andersen K, Bird KL, Rasmussen M, Haile J, Breuning-Madsen H, Kjær KH, Orlando L, Gilbert MTP, Willerslev E. Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. Mol Ecol 2012; 21:1966-1979.
10. Taberlet P, Prud'Homme SM, Campione E, Roy J, Miquel C, Shehzad W, Gielly L, Rioux D, Choler P, ClÉMent J-C *et al*. Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Mol Ecol 2012; 21:1816-1820.
11. Ficetola GF, Miaud C, Pompanon F, Taberlet P. Species detection using environmental DNA from water samples. Biology Letters 2008; 4:423-425.
12. Thomsen PF, Kielgast JOS, Iversen LL, Wiuf C, Rasmussen M, Gilbert MTP, Orlando L, Willerslev E. Monitoring endangered freshwater biodiversity using environmental DNA. Mol Ecol 2012; 21:2565-2573.
13. Shokralla S, Spall JL, Gibson JF, Hajibabaei M. Next-generation sequencing technologies for environmental DNA research. Mol Ecol 2012; 21:1794-1805.
14. Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH. Environmental DNA. Mol Ecol 2012; 21:1789-1793.
15. Brunbjerg AK, Bruun HH, Moeslund JE, Sadler JP, Svenning J-C, Ejrnæs R. Ecospace: A unified framework for understanding variation in terrestrial biodiversity. Basic and Applied Ecology 2017; 18:86-94.
16. Belyea LR, Lancaster J. Assembly rules within a contingent ecology. Oikos 1999; 86:402-416.
17. Ricklefs RE. Community diversity - relative roles of local and regional processes. Science 1987; 235:167-171.
18. Horák J, Kout J, Vodka Š, Donato DC. Dead wood dependent organisms in one of the oldest protected forests of Europe: Investigating the contrasting effects of within-stand variation in a highly diversified environment. Forest Ecology and Management 2016; 363:229-236.
19. Seibold S, Bässler C, Brandl R, Gossner MM, Thorn S, Ulyshen MD, Müller J. Experimental studies of dead-wood biodiversity — a review identifying global gaps in knowledge. Biological Conservation 2015; 191:139-149.
20. Laliberté E, Zemunik G, Turner BL. Environmental filtering explains variation in plant diversity along resource gradients. Science 2014; 345:1602-1605.

21. Houseman GR, Gross KL. Linking grassland plant diversity to species pools, sorting and plant traits. Journal of Ecology 2011; 99:464-472.
22. Rubæk GH, Kristensen K, Olesen SE, Østergaard HS, Heckrath G. Phosphorus accumulation and spatial distribution in agricultural soils in Denmark. Geoderma 2013; 209–210:241-250.
23. Nygaard B, Juel A, Fredshavn JR. Ændringer i det § 3-beskyttede naturareal 1995-2014. Resultater fra Naturstyrelsens opdateringsprojekt. Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi, 106 s. - Teknisk rapport fra DCE - Nationalt Center for Miljø og Energi nr. 79. http://dce2.au.dk/pub/TR79.pdf; 2016.
24. Arler F, Jørgensen MS, Galland D, Sørensen EM. Kampen om m^2 - Prioritering af fremtidens arealanvendelse i Danmark. Fonden Teknologirådet. http://www.tekno.dk/wp-content/uploads/2015/08/Prioritering-af-fremtidens-arealanvendelse-i-Danmark.pdf; 2015.
25. Terrestriske Naturtyper 2004 – 2015. NOVANA. Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi. www.novana.au.dk.
26. Council Directive 92/43/EEC. On the conservation of natural habitats and of wild fauna and flora. European Commission. 1992.
27. Hill MO, Mountford JO, Roy DB, Bunce RGH. Ellenberg's indicator values for British plants : ECOFACT volume 2 : Technical Annex. Natural Environment Research Council: Centre for Ecology and Hydrology; 1999.
28. Grime JP, Hodgson JG, Hunt R. Comparative plant ecology: a functional approach to common British species. London: Unwin Hyman; 1989.
29. Ejrnæs R, Bruun HH. Gradient analysis of dry grassland vegetation in Denmark. Journal of Vegetation Science 2000; 11:573-584.
30. Lomolino MV, Riddle BR, Brown JH, Brown JH. Biogeography. In.: Sinauer Associates Sunderland, MA; 2006: 65-96.
31. Bijl Lvd, Boutrup S, Jensen PN. NOVANA. Det nationale program for overvågning af vandmiljøet og naturen. Programbeskrivelse 2007-2009 - del 2. Danmarks Miljøundersøgelser, Aarhus Universitet. Faglig rapport fra DMU nr 615 2007.
32. Schmidt P-A, Bálint M, Greshake B, Bandow C, Römbke J, Schmitt I. Illumina metabarcoding of a soil fungal community. Soil Biology and Biochemistry 2013; 65:128-132.
33. White N, Engeman R, Sugihara R, Krupa H. A comparison of plotless density estimators using Monte Carlo simulation on totally enumerated field data sets. BMC Ecology 2008; 8:6.
34. iNEXT: iNterpolation and EXTrapolation for species diversity. R package version 2.0.12 URL: http://chao.stat.nthu.edu.tw/blog/software-download/.
35. R Core team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2015.
36. Oksanen J, Blanchet FG, Kindt R, Legendre P, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. Package 'vegan': Community Ecology Package. Version 1.17-2. http://cran.r-project.org/web/packages/vegan/vegan.pdf. 2010.
37. Kraft NJB, Adler PB, Godoy O, James EC, Fuller S, Levine JM. Community assembly, coexistence and the environmental filtering metaphor. Functional Ecology 2015; 29:592-599.
38. Gillison AN, Brewer KRW. The use of gradient directed transects or gradsects in natural resource surveys. Journal of Environmental Management 1985; 20:103-127.
39. Standen V. The adequacy of collecting techniques for estimating species richness of grassland invertebrates. Journal of Applied Ecology 2000; 37:884-893.
40. Popic TJ, Davila YC, Wardle GM. Evaluation of common methods for sampling invertebrate pollinator assemblages: net sampling out-perform pan traps. PLoS One 2013; 8:e66665.

624 41. CaraDonna PJ, Petry WK, Brennan RM, Cunningham JL, Bronstein JL, Waser NM, Sanders NJ.
625 Interaction rewiring and the rapid turnover of plant–pollinator networks. Ecology Letters 2017;
626 20:385-394.
627 42. Valverde J, Gómez JM, Perfectti F. The temporal dimension in individual-based plant pollination
628 networks. Oikos 2016; 125:468-479.
629 43. Brown SP, Veach AM, Rigdon-Huss AR, Grond K, Lickteig SK, Lothamer K, Oliver AK, Jumpponen
630 A. Scraping the bottom of the barrel: are rare high throughput sequences artifacts? Fungal
631 Ecology 2015; 13:221-225.
632 44. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. Wrinkles in the rare biosphere:
633 pyrosequencing errors can lead to artificial inflation of diversity estimates. Environmental
634 Microbiology 2010; 12:118-123.
635 45. Bálint M, Bahram M, Eren AM, Faust K, Fuhrman JA, Lindahl B, O'Hara RB, Öpik M, Sogin ML,
636 Unterseher M *et al*. Millions of reads, thousands of taxa: microbial community structure and
637 associations analyzed via marker genes. FEMS Microbiology Reviews 2016; 40:686-700.

638