Protistan community analysis: key findings of a large-scale molecular sampling

Lars Grossmann¹, Manfred Jensen¹, Dominik Heider², Steffen Jost¹, Edvard Glücksman³, Hanna Hartikainen⁴,⁵, Shazia S Mahamdallie⁴,⁶, Michelle Gardner⁴, Daniel Hoffmann⁷, David Bass⁴,⁸ and Jens Boenigk¹

¹Biodiversity Department, Centre for Water and Environmental Research, University of Duisburg-Essen, Essen, Germany; ²Department of Bioinformatics, Straubing Centre of Science, Straubing, Germany; ³Environment & Sustainability Institute, University of Exeter, Penryn, Cornwall, UK; ⁴Department of Life Sciences, The Natural History Museum, London, UK; ⁵Eawag and Institute for Integrative Biology, ETH Zurich, Duebendorf, Switzerland; ⁶Division of Genetics & Epidemiology, Institute of Cancer Research, London, UK; ⁷Bioinformatics Department and Centre for Water and Environmental Research, University of Duisburg-Essen, Essen, Germany and ⁸Centre for Environment, Fisheries and Aquaculture Science (Cefas), The Nothe, Weymouth, Dorset, UK

Protists are perhaps the most lineage-rich of microbial lifeforms, but remain largely unknown. High-throughput sequencing technologies provide opportunities to screen whole habitats in depth and enable detailed comparisons of different habitats to measure, compare and map protistan diversity. Such comparisons are often limited by low sample numbers within single studies and a lack of standardisation between studies. Here, we analysed 232 samples from 10 sampling campaigns using a standardised PCR protocol and bioinformatics pipeline. We show that protistan community patterns are highly consistent within habitat types and geographic regions, provided that sample processing is standardised. Community profiles are only weakly affected by fluctuations of the abundances of the most abundant taxa and, therefore, provide a sound basis for habitat comparison beyond random short-term fluctuations in the community composition. Further, we provide evidence that distribution patterns are not solely resulting from random processes. Distinct habitat types and distinct taxonomic groups are dominated by taxa with distinct distribution patterns that reflect their ecology with respect to dispersal and habitat colonisation. However, there is no systematic shift of the distribution pattern with taxon abundance.

Introduction

Protists are abundant and diverse in aquatic and terrestrial ecosystems and fulfil critical ecosystem functions (del Campo and Massana, 2011; Triadó-Margarit and Casamayor, 2012; Bates et al., 2013; Geisen et al., 2015). They not only strongly contribute to primary production (Field et al., 1998) and bacterial grazing (Boenigk and Arndt, 2002; Glücksman et al., 2010), but are also major players in diverse nutrient cycles (Finlay and Esteban, 1998; Coleman and Whitman, 2004) and as such, they form the basis of aquatic and terrestrial food webs along with bacteria.

The ecological key role of protists is sharply contrasted by an only vague understanding of their diversity. During the past decade, the assumption of unlimited dispersal of protists was progressively replaced by the ‘moderate endemicity model’ (Foissner, 2006; Bass and Boenigk, 2011; but see Livermore and Jones, 2015). With the increasing acceptance of dispersal limitations in protists, the research focus turned to the underlying patterns of protist distribution and the generalisability of such patterns across protistan taxa and across different habitat types (Katz et al., 2005; Bik et al., 2012). Owing to methodical limitations, analyses of protist distribution patterns up to recently were centred around a handful of more or less abundant taxa (for example, Lopéz-García et al., 2001; Moreira and López-García, 2002). High-throughput sequencing has already revolutionised environmental surveys of microbial organisms and opened the door for large-scale analyses of molecular microbial community analyses. Generating many millions of reads, these methods allow for deep-sequencing of microbial communities (Caporaso et al., 2012) and make possible a much more realistic understanding of protist diversity and distribution than former, lower capacity technologies (Medinger et al., 2010; Lecroq et al., 2011; Degnan and Ochman, 2012).
The current high-throughput sequencing technologies now allow for sufficiently deep community analysis. Nevertheless, most studies have been restricted to few sampling sites (Caron and Countway, 2009; Stoeck et al., 2009; Nolte et al., 2010; Caporaso et al., 2012). Only recently, large-scale multisample sets, specifically from the marine biome (de Vargas et al., 2015), allow for deeper insights into protist distribution patterns. Here, we address the generalizability of protist distribution patterns with respect to habitat type, taxonomic group and taxon abundance based on a large dataset comprising 232 samples from soil as well as fresh, brackish and marine waters. Besides judging the relative importance of habitat type, seasonality and geographic region in shaping protist distribution patterns, we evaluate the question on whether or not rare taxa show similar distribution patterns. For analyses of higher taxonomic group level, the assembled data set are summarised in Supplementary Table S1. The 10 different sampling campaigns representing eight different habitat types are referred to as Alpine spatial, Alpine seasonal, Lake District, Winterpico, Estuarine, Soil, Whale Fall, Borehole, heterotrophic nanoflagellates (HNF) and Biofilm within this paper. Briefly, they comprise an alpine transect of freshwater samples (Austria, 32 lakes at 450–2050 m elevation) and a seasonal sampling of three Austrian lakes (Lake Mondsee region, 9–28 samples per lake, range: April to December), a seasonal sampling in the English Lake District, partially size-fractionated (UK, 7 sites, 2–3 samples per site) and size-fractionated samples of the same sites in winter (Winterpico), an estuarine gradient (UK, River Colne, 11 sites, partially seasonal), a comparative soil sampling from fallow and cultivated fallow fields (UK, 15 sites across England), a sampling on the seabed of and from below a sunken whale (Whale fall) (North Sea, Sweden), a collection of aquifer samples from boreholes (UK, Berkshire), an experimental set-up of a size-filtered (5 μm) freshwater sample fed with different food bacteria and focusing on heterotrophic nanoflagellates (Czech Republic, Křímov reservoir) and an experimental set-up of freshwater biofilms treated with the viricide TamiFlu (Southern UK). Owing to the different natures of the collected samples, slightly modified sampling protocols and DNA-isolation methods (compared in Supplementary Table S1) had to be used.

Materials and methods

Sampling and sample preparation

Sampling sites and sampling procedures of the assembled data set are summarised in Supplementary Table S1. The 10 different sampling campaigns representing eight different habitat types are referred to as Alpine spatial, Alpine seasonal, Lake District, Winterpico, Estuarine, Soil, Whale Fall, Borehole, heterotrophic nanoflagellates (HNF) and Biofilm within this paper. Briefly, they comprise an alpine transect of freshwater samples (Austria, 32 lakes at 450–2050 m elevation) and a seasonal sampling of three Austrian lakes (Lake Mondsee region, 9–28 samples per lake, range: April to December), a seasonal sampling in the English Lake District, partially size-fractionated (UK, 7 sites, 2–3 samples per site) and size-fractionated samples of the same sites in winter (Winterpico), an estuarine gradient (UK, River Colne, 11 sites, partially seasonal), a comparative soil sampling from fallow and cultivated fallow fields (UK, 15 sites across England), a sampling on the seabed of and from below a sunken whale (Whale fall) (North Sea, Sweden), a collection of aquifer samples from boreholes (UK, Berkshire), an experimental set-up of a size-filtered (5 μm) freshwater sample fed with different food bacteria and focusing on heterotrophic nanoflagellates (Czech Republic, Křímov reservoir) and an experimental set-up of freshwater biofilms treated with the viricide TamiFlu (Southern UK). Owing to the different natures of the collected samples, slightly modified sampling protocols and DNA-isolation methods (compared in Supplementary Table S1) had to be used.

PCR and pyrosequencing

PCR amplifications targeting the SSU V9 region were conducted using primers 1391F (Lane, 1991; Stoeck et al., 2010) and EukB (Medlin et al., 1998; Stoeck et al., 2010) both carrying a 5′-tail for the 454 sequencing (adapter A: GCCTCCCTCGGCCTCATCAG, adapter B: GCCTTGGAGCGCCCTTCAG) to amplify a broad spectrum of eukaryotes. The final concentrations in all of the PCR reactions were as follows: 1 μl of DNA template in 20 μl PCR reaction with 0.4 units of Phusion polymerase; primers at 0.25 μM final concentration; dNTPs at 0.2 mM final concentration; 4 μl Phusion buffer; and 12.2 μl water. The PCR conditions consisted of an initial denaturation at 94 °C for 4 min and 35 cycles of: 30 s at 95 °C, annealing for 60 s at 60 °C, elongation for 2 min at 72 °C, followed by a final extension step of 10 min at 72 °C. We received sequences of about 200bp from PCR. Pyrosequencing was carried out on a 454 Genome Sequencer FLX System using the Titanium chemistry (454 Life Science, Branford, CT, USA) with 50–100 amplified samples placed onto one 454 sequencing platform, respectively.

Bioinformatics

For the bioinformatics analyses, we used a standardised pipeline including (i) quality filtering, (ii) clustering and (iii) taxonomy annotation. Low-quality tails were removed from the reads and trimmed reads with an average Phred quality score less than 25 were discarded. Additionally, we removed all reads with at least one base with a quality of less than 15 and also all reads that contained errors in the primer regions. Multiplex Identifiers were used to separate the different samples. Chimeras were removed using UCHIME (USEARCH v.6) (Edgar et al., 2011). Subsequently, the sequences that passed the quality filtering were clustered into Operational Taxonomic Units (OTUs) with UCLUST version 6 (Edgar, 2010) at 97% sequence identity for further analyses. Sequences were sorted by abundance beforehand clustering and clustering was performed for all samples at a time (that is, not individually by campaigns). For all OTUs, we used BLASTn version 2.2.25+ (Altschul et al., 1990) with the nucleotide collection database nr/nt and the NCBI Taxonomy Database (http://www.ncbi.nlm.nih.gov/taxonomy; RefSeq Release 57 (14 January 2013)) to annotate the OTUs with taxonomic information. For reasons of comparison, we also processed our data with the CANGS pipeline (Pandey et al., 2010) and received similar results.

Taxonomic analysis of the sequence data set

OTUs and their reads from bioinformatic filtering and assembling were used to analyse the community structure and distribution patterns of OTUs (abundant vs rare, habitat-wise, and taxon-wise). For analyses on higher taxonomic group level, the following taxa were selected: Ciliophora, Dinophyceae, Apicomplexa, Alveolata rest, Bicosoecida,
Oomycetes, Chrysophyceae, Synurophyceae, Bacil-
leriophyta, Stramenopiles rest, Cercozoa, Rhizaria
rest, Heterolobosea, Euglenida, Kinetoplastida,
Euglenozoa rest, Excavata rest, Choanoflagellida,
Chytridiomycota, Microsporidia, Ascomycota, Basi-
diomyctota, Glomeromycota, Amoebozoa, Apusozoa,
Glaucocystophyceae, Rhodophyta, Viridiplantae
without Embryophytes, Cryptophyta, Katablephari-
dophyta and Haptophyta. OTUs affiliated with
metazoa and embryophytes (as multicellular organ-
isms), as well as with Bacteria and Archaea, have
been excluded before further analyses. Thus, only
protists were analysed since they were the focus of
our investigation. The ‘rest’ categories within the
analysis are meant to collect minor groups and OTUs
without further affiliation within Alveolata, Strame-
nopiles, Rhizaria and Excavata in order to fully show
the protist diversity present in the samples.
Eukaryotic OTUs that are not met by one of the
given taxonomic groups or have not been deleted from
the data set for the given reasons are shown as
‘others’. Furthermore, OTUs that could not be
affiliated to entries in the NCBI database are labelled
as ‘unknown’.

Statistical analysis of the sequence data set
The specific nature of the data matrix that results
from deep-sequencing diversity counting is charac-
terised by a multitude of zero and single counts of
OTUs (zero: 98.35%, single counts: 0.83%, two
counts: 0.27%, ten counts: 0.014%, etc.) not fitting
traditional statistical methods of diversity com-
parison. Therefore, we used a two-step standardisation
of site data; first, rarefaction of the sites was done
using the ‘drarefy’ function provided by the VEGAN
R package (Oksanen et al., 2013), followed by
Hellingrier transformation (Legendre and Gallagher,
2001). After application of these procedures, the
resulting pretransformed data matrix of OTUs can be
submitted to Ward cluster analysis and further multivariate analyses (see Figures 1a and b;
Legendre and Gallagher, 2001; Borcard et al.,
2011). Cluster analysis could be calculated on both
the OTU and the meta-group level (for chosen meta-
groups see above). The principle component analysis
based on Hellingier distances shows those taxonomic
groups with most explanatory value for clustering
and separation of samples (see Figure 1b).

As reads per sample differ decisively between
samples, rarefaction values at a read level of 100
reads (including extrapolation for a few samples, see
Colwell et al., 2012) were used to compare species
richness and composition of OTUs among samples.
For the shown composition within each sample, the
‘drarefy’ function (R package VEGAN; Oksanen et al.,
2013) was used to minimise the bias in the rarefac-
tion analysis. The more classical ‘rarefy’ is not
suitable here, as recently pointed out by McMurdie
and Holmes (2014). Thus, more exact relationships
of meta-groups from higher read levels are kept
intact (see Figure 1a); however, this could possibly
result in OTU percentages per meta-group of less
than 1 at a read level of 100. The community struc-
tures of the different samples were analysed using
the mean relative abundance of meta-group reads per
sampling, which showed significant differences
between meta-groups among sampling campaigns
(see Figure 2, also compare Supplementary Table S2).

Weibull distribution in protistan OTUs
Following preliminary tests, the distribution
patterns of individual OTUs did not follow a Poisson
distribution (P<0.05) or a normal distribution (Ricci,
2005). Although not common, a two-parameter
Weibull density distribution, which is used in
survival and engineering analyses, fitted our data.
Recently, a three-parameter Weibull empirical
cumulative distribution function was used for fitting
diversity curves (Livermore and Jones, 2015), showing
the applicability of the Weibull distribution in
ecology. For our data, the fitting of the Weibull
distribution among sites (y-axis=no. of sites, x-axis=
abundance of OTU in these sites) passed the
KS-test (that is, the null hypothesis could not be
rejected, P-value: >0.05) for almost all OTUs and
was consequently used as the distribution for further
analysis. All KS-tests were performed with R,
referring to Birnbaum and Tingey (1951). All OTUs
passed the KS-test (that is, the Weibull distribution
could not be rejected with all P-values >0.9) when
P-values were corrected for multiple testing by using
p.adjust (method = ‘fdr’, R package ‘stats’). When
corrections for multiple testing were disregarded,
the KS-test rejected the Weibull distribution for 10 out
of 1249 OTUs. Even though these 10 OTUs passed
the test when correctly applied (that is, corrected for
multiple testing), we decided to mark these OTUs in
the figures (Figures 3a and b, top). The shape
density of the Weibull distribution is suitable
for comparing the distribution of abundant and less
abundant taxa as the total abundance primarily
affects the scale, but not the shape parameter.
Therefore, we focus on the comparative analysis of
the parameter k. For k=1, the two-parameter
Weibull distribution switches to that of the exponen-
tial distribution, for k ~3.6 it switches to a
Gaussian distribution. The k-values in this study all
lie between 1 and 3.6 forming curves of positive
skewness. The skewness of the curves as a function
of k indicates the value of evenness of OTU reads
over sites and is, therefore, suitable of showing
distribution patterns of taxa. Shape values k of
Weibull distributions were compared using Mann–
Whitney U-tests. All U-tests were performed with
Sigmamplot 12.5 (Systat Software Inc., Erkrath,
Germany) referring to Mann and Whitney (1947).
The control data set, tested against Bacilliariophyta as
the closest subgroup data set within the analysed
data, was generated by the function ‘rnorm’ with
Figure 1  Results of cluster analysis of sites. (a) Bars (left to right) are ordered by clusters from cluster analysis which almost entirely fit the actual sampling campaigns (see cluster names); black arrows at bottom indicate those sites within clusters not identified with the respective sampling campaign; within clusters, order (left to right) is by species richness of identified taxonomic groups, respectively; species richness is shown as height of bars out of 100; community composition is shown as colour of bars by identified taxonomic groups (bottom) and merged or unidentified reads (top) (procedure: drarefy, see Materials and methods). (b) PCA showing clusters from cluster analysis (encircled) and significant taxonomic groups for sorting of sites (arrows) (pretreatment of data: drarefy+Hellinger transformation, see Materials and methods); PCA was calculated on the basis of percentage of taxonomic groups within sites. Abbreviation: PCA, principle component analysis.
n = 55 (normally distributed), average value and median = 1 (Ricci, 2005), and with the same standard deviation value as calculated for the Bacillariophyta data set. We also provide a graph with exemplary Weibull curves of typical OTUs in the Supplementary Information on methods.

Results
The large-scale analysis undertaken herein—carried out as a multiple sample habitat comparison—showed that protistan communities differ decisively. They did so in terms of OTU presences in a distinct sample and habitat as well as in the ratio of higher taxonomic groups (meta-groups) (Figure 1a). Taking all samples into account, Ciliophora were dominating in terms of d-rarefied sequence abundance at 13% followed by Bacillariophyta at 9.7% and Chrysophyceae at 9%.

Community composition patterns
Despite many OTUs being distributed across different campaigns and sample types, many OTUs were either specific for distinct habitats/campaigns or their relative abundance deviated considerably from that in other habitats/campaigns. Cluster analysis based on protistan sequence analysis separated different habitat types. Specifically soil, biofilm and brackish sediments (estuarine), and freshwaters were distinguished and each showed a specific community composition (Figure 1b). Size-fractionated samples, such as the heterotrophic nanoflagellates data set, were also separated by cluster analysis.

This differentiation based on OTUs was also reflected by shifts in the community composition on the level of higher taxonomic groups (Figures 1b and 2). Substrate-bound taxa were, as expected, specifically abundant in sediments. However, while Amoebozoa and Cercozoa were specifically more abundant in the soil samples (see column A in Figure 2), Heterolobosea, Kinetoplastida, Apusozoa and to some extent Chrysophyceae were specifically abundant in the aquifers of the borehole samples (see column B in Figure 2). The Peridiosporomycota (Oomycetes) were found to be relatively abundant in both habitats. In the size-fractionated samples, small heterotrophic nanoflagellates taxa showed relatively high sequence abundances, specifically Chrysophyceae, Kataablepharidophyceae, Choanomonada and Bicosocida (see column F in Figure 2). In our biofilm samples, we found an increased sequence abundance of Chlorophyta and streptophytic algae.
Protistan community analysis
L. Grossmann et al.

(a) Weibull shape by habitats

Eukaryota all

Alpine Lakes and Lake District I

Soil

Weibull shape

no. of OTUs (with sequence OTU)

no. of OTUs (without sequence OTU)

log abundance of OTUs

(b) Occurrence of OTUs in sites and Weibull shape by taxonomic groups

Eukaryota all

O. TUs present in many sites few sites

Alveolata

Amoebozoa

Chrysophyceae

Bacillariophyceae

Weibull shape

log abundance of OTUs

The ISME Journal
(labelled as ‘Viridiplantae’ without Embryophytes), Cryptophyta, Bacillariophyta and Chytridiomycota. Substrate-bound taxa did not have particularly high sequence abundances in the biofilm samples (see column C in Figure 2). The marine samples showed comparatively low read counts of Chlorophyta, streptophytic algae, Chytridiomycota and Cryptophyta, but comparatively high reads of Bacillariophyta in the salinity gradient along the Colne estuary, and of Dinophyceae and Euglenida at the whale fall site (see column D in Figure 2).

These habitat-specific community profiles were similar for almost all samples of the respective habitat types. Cluster analysis revealed only very few samples that would be sorted into the wrong campaign based on sequence data (see arrows in Figure 1a).

**Differential distribution patterns of protistan taxa**

For the analysis of OTU distribution patterns among sites, we analysed the shape parameter \( k \) of the Weibull distribution (see the Materials and methods section). This analysis only revealed a minor modulation in the shape factor of the Weibull distribution between abundant and less abundant taxa. The mean Weibull shape value \( k \) is rather similar irrespective of the relative overall abundance of the OTUs (Figure 3a, top). Shape values of individual OTUs scatter around a median of 1.67, but there was no distinct change in the median \( k \) for less abundant OTUs. The median shape value of 1.67 indicates that taxon distribution is not due to unlimited random dispersal (which would favour a value of 1 or lower, that is, a colonisation from one site with high frequency), but on the contrary that some factors work against invasion and/or colonisation of habitats (significant \( U \)-test of shape value 1.67 vs 1, \( P \)-value: <0.001). One might argue that this effect is caused by the different habitat types included in the analysis. However, even when the analysis is restricted to one habitat type within a distinct geographic region, taxon distribution cannot exclusively be explained by unlimited random dispersal (for example, soil sites UK: shape parameter = 2.099, significant \( U \)-test vs 1, \( P \)-value: <0.001; for example, Austrian lakes+Lake District 1: shape parameter = 1.711, significant \( U \)-test vs 1, \( P \)-value <0.001; compare methods in this paper and Figure 3a—mid and bottom). In contrast, the shape parameter is even higher when the analysis is restricted to soils, indicating that random dispersal is even less important in soils compared with aquatic habitats. However, the scatter around the median \( k \) increases for rare taxa (Figure 3b, top right). Similarly, the scatter of sites of occurrence as a function of total read number increases (Figure 3b, top left). In contrast, the most abundant taxa do have very similar distribution patterns, indicated by both a very similar shape parameter in the Weibull distribution (Figure 3b, top right) and a narrow range of the ratio of number of sites vs total read abundance (Figure 3b, top left). Both observations indicate a stronger differentiation of distribution patterns in rare taxa, whereas abundant taxa show relatively uniform distribution patterns. Owing to this, within the rare taxa, a differentiation into more generalist taxa (occurring at a relatively high number of sites) and more specialised taxa (restricted to a few sites) is pronounced.

Similarly, distribution differs between taxonomic groups. Diatom OTUs, as one extreme, are characterised by a comparatively low \( k \) and occur at a relatively low number of sites for a given total read count. Amoebozoa, on the other hand, are characterised by comparatively high shape values and occur at a relatively high number of sites for a given total read count. The distribution pattern of Amoebozoa indicates a broader (more generalist) distribution, whereas the Diatom distribution pattern reflects a narrower (more specialised) distribution (significant \( U \)-test of shape values for Amoebozoa vs Bacillariophyceae, \( P \)-value: <0.001). The Chrysophyceae are somewhat in between the two strategies. Nevertheless, \( k \) is significantly larger as would be expected for random distribution for all taxa (Amoebozoa: shape parameter = 2.139, significant \( U \)-test vs 1, \( P \)-value: <0.001; Bacillariophyceae: shape parameter = 1.449, significant \( U \)-test vs 1, \( P \)-value: <0.001; Chrysophyceae: shape = 1.595,
significant $U$-test vs 1, $P$-value: 0.009) (Figure 3b, right and left)

Non-random distribution patterns are also supported by principle component analysis (Figure 1b). Based on the molecular diversity, the samples clustered according to habitat types. Nevertheless, within habitat types, regional effects were pronounced. Samples from the same sampling site clustered together irrespective of seasonal effects. Furthermore, samples from the same region (for example, from Austria) mostly clustered together.

**Discussion**

Here, we address the generalisability of protist distribution patterns. We demonstrate that protist taxon distribution varies depending on habitat types, taxonomic groups and taxon abundance.

With our analysis, we offer a large-scale and cross-biome investigation of protist diversity and distribution. The standardised comparison across diverse habitats is, to our knowledge, unique to this study and includes standardised nucleic acid processing, sequencing, bioinformatics, and statistical analyses of sediment, biofilm and water samples of diverse locations. Owing to the approach of comparing different habitat types, sampling methods had to be altered. Standardisation is otherwise complete and, therein, goes beyond comparative platform approaches (Caporaso et al., 2012; Mahé et al., 2014). Further, even though an unbiased view on protist diversity is difficult to achieve (compare Supplementary Information on methods), the standardisation over all 232 samples of the analysis allows for addressing overarching patterns of protist diversity in a comparative approach.

**General patterns of dispersal limitation**

Protist distribution and potential dispersal limitations have been in the focus of protist ecology and diversity research since the early observations. The scientific dispute can be traced back at least to de Candolle (1820) and Ehrenberg (1849, 1850). With the catchy phrase ‘Everything is everywhere, but the habitat selects’ Baas-Becking formulated the central idea behind the ongoing dispute (Bass and Boenigk, 2011; Boenigk, 2014). In this form, the idea of global dispersal was revived in the 1980s, resulting in the recent debate on protist dispersal and endemism as summarised by O’Malley (2007), de Wit and Bouvier (2006) and Bass and Boenigk (2011). It is now generally believed that protist distributions cannot be explained exclusively by unlimited dispersal (Foissner, 2006; Bass and Boenigk, 2011), expressed, for instance, by the moderate endemism model (Foissner, 1999, 2004, 2006). However, microbial organisms may, in general, show wider distribution patterns as compared with higher organisms (Livermore and Jones, 2015). We demonstrate that the distribution patterns of individual protistan OTUs, specifically as expressed by the shape parameter of the Weibull distribution, violate the assumption of unlimited dispersal. This is consistent with findings of Livermore and Jones (2015) who demonstrated for bacteria that microbial diversity patterns non-randomly vary between ecosystem types. Our analysis further shows that patterns of non-random dispersal are stronger in soil habitats than aquatic habitats.

**Consistency of distribution patterns within habitat types**

Molecular data produce strong protistan community signatures, that is, habitats being characterised by distinct protistan communities (Bates et al., 2011; Tedersoo et al., 2014; de Vargas et al., 2015). Consistent with earlier findings from microscopy and Sanger sequencing studies (Fenchel, 1994; Anderson, 2012; Tikhonenkov et al., 2012), soil and aquifer protist communities are dominated by substrate-bound taxa (Foissner, 1991; Novarino et al., 1997; Adl, 2003; Adl and Coleman, 2005; Risse-Buhl et al., 2013). Likewise, the marine samples showed a high proportion of sequences affiliated with diatoms, dinoflagellates and euglenids (Ingmanson and Wallace, 1995; Graham et al., 2009), whereas Synurales, Cryptophytes, Chlorophyta and Chytridiomycota were particularly prominent in freshwater samples (Dokulil et al., 2001; Graham et al., 2009).

The consistency of distribution patterns with respect to habitat type was also reflected at the level of individual OTUs. Principle component analysis, as well as cluster analysis, revealed similar community profiles for samples originating from similar habitats. These community profiles were highly specific for distinct habitat types. Even seasonal variation did not blur the distinctiveness of community structure. Therefore, community profiles based on high-throughput sequencing data (in this case the 18S rDNA V9 region) appear highly promising for overarching habitat comparisons as well as for long-term monitoring studies.

**The relatively uniform distribution pattern of abundant taxa is contrasted by a high variation in rare taxa**

Molecular studies typically report a high fraction of rare taxa. The high number of rare taxa led to the assumption that rarity might be an evolutionarily advantageous trait in these organisms (Logares et al., 2014) and favoured speculations on potentially deviating distribution patterns between rare and abundant taxa (for example, Nolte et al., 2010). In contrast to these earlier studies, we show that rare and abundant taxa have generally similar distribution patterns. However, whereas the distribution patterns of individual taxa are relatively uniform for abundant OTUs, the deviation among individual rare OTUs is high. Further, the proportion of rare and
abundant taxa differs distinctively by habitat type: the soil samples are characterised by a high evenness, that is, a multitude of OTUs appear at higher abundances in many of the samples. On the contrary, aquatic sites have a comparatively lower evenness with usually only a small number of dominant species reaching higher abundances, accompanied by many rare taxa, at a distinct point in time. Thus, in addition to taxonomic profile/community structure differences between habitats, abundance and diversity patterns appear to be characteristic for different habitat types.

Differences in distribution patterns indicate differential niche widths between taxa (Boenigk et al., 2006; Bass et al., 2009). In general, broad niche widths and high ecophysiological tolerances should be reflected by broad distributions of taxa across different samples. In contrast, taxa with more restricted niche widths and/or ecophysiological tolerances should be restricted to fewer samples. The largely uniform distribution patterns for abundant OTUs indicate that dispersal and success of colonisation is largely comparable between taxa, resulting in similar distribution patterns. Within the rare biosphere, however, distribution patterns vary strongly between individual taxa, presumably indicating a more selective niche adaptation. Thus, we suspect that the rare biosphere offers an even stronger potential use for comparative habitat analyses, including biological monitoring, as compared to the more abundant taxa, which are usually the focus of such studies.

**Differential patterns between taxonomic groups**

With respect to taxonomic groups, amoebozoan taxa (that is, amoebozoan OTUs in our analysis) are often characterised by broader distribution patterns, indicating a more generalist strategy, whereas the diatom taxa (that is, OTUs) are often characterised by narrower distribution patterns, indicating a more specialised strategy. The Chrysophytes, for instance, show intermediate distribution patterns. These differences in distribution presumably indicate a differential dispersal and proliferation among taxa even on higher taxonomic levels, including not all, but most OTUs of a higher taxonomic group. With regard to taxonomy, it seems plausible that distributions are markedly different between soil-inhabiting and endobenthic taxa, such as Amoebozoa, which are presumably less prone to random dispersal, and planktonic or epibenthic taxa.

**Conclusions**

With our analyses, we show that limited dispersal and distribution in protists differ by habitat type as well as by taxonomic group. Further, rare and abundant taxa do not show generally different patterns of distribution. The variation of distribution patterns in rare taxa is, however, much more pronounced making the rare biosphere as revealed by molecular investigations a promising candidate for comparative habitat analyses. For analyses as undertaken herein, standardisation of a large number of samples is essential and can be provided by large-scale, methodologically unified data sets and parallel investigations made possible by high-throughput methods. Such investigations are more directly comparable than data sets generated by different ‘wet’ laboratory conditions (PCR protocols, cycling conditions, reagents, etc.) and different bioinformatic methods. This unified/standardised approach is recommended for revealing robust and fine-scale differences between community structures and distributions of taxa that might be missed or skewed by non-standardised comparisons. This is important for further illuminating the biodiversity and biology of protists on Earth and thus broaden our knowledge on these and other under-researched microorganisms.

**Conflict of Interest**

The authors declare no conflict of interest.

**Acknowledgements**

We thank all cooperation partners, especially Thomas Bell, Gary Bending, Adrian Glover, Paul Gosling, Rene Groben, Martin Hahn, Sally Hilton, Katja Lehmann, Stephen Maberly, Cecilia Martinez-Perez, Louise Maurice, Mark Osborn, Kevin Purdy, Karel Šimek, Andrew Singer, Cindy Smith, Peter Studer and CEH Lancaster and Wallingford (UK) for access to samples and assistance with sampling. We thank the German Research Foundation (BO-3245/2) for financial support.

**References**

Adl MS. (2003). *The Ecology of Soil Decomposition*. CABI Publishing: Wallingford, UK.

Adl MS, Coleman DC. (2005). Dynamics of soil protozoa using a direct count method. *Biol Fertil Soils* 42: 168–171.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* 215: 403–410.

Anderson OR. (2012). The role of bacterial-based protist communities in aquatic and soil ecosystems and the carbon biogeochemical cycle, with emphasis on naked amoebae. *Acta Protozool* 51: 209–221.

Bass D, Boenigk J. (2011). Everything is everywhere: a twenty-first century de/reconstruction with respect to protists. In: Fontaneto D (ed), *Biogeography of Microscopic Organisms: Is Everything Small Everywhere?* CUP: Cambridge, pp 88–110.

Bass D, Brown N, Mackenzie-Dodds J, Dyal P, Nierzwicki-Bauer SA, Vepritskiy AA et al. (2009). A molecular perspective on ecological differentiation and
biogeography of cyclotrichiid ciliates. *J Eukaryot Microbiol* 6: 559–567.

Bates ST, Berg-Lyons D, Caporaso JG, Walter WA, Knight R, Fierer N. (2011). Examining the global distribution of dominant archeal populations in soil. *ISME J* 5: 908–917.

Bates ST, Clemente JC, Flores GE, Walters WA, Wegener Parfrey L, Knight R et al. (2013). Global biogeography of highly diverse protistan communities in soil. *ISME J* 7: 652–659.

Bik HM, Sung W, De Ley P, Baldwin JC, Sharma J, Rocha-Olivares A et al. (2012). Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Mol Ecol* 21: 1048–1059.

Birnbaum ZW, Tingey FH. (1951). One-sided contours for probability distribution functions. *Ann Math Stat* 22: 592–596.

Boenigk J. (2014). Five decades of research in protozoology – what have we learned? In: Hausmann K, Radek R (eds), *Cilia/Flagella—Ciliates/Flagellates*. Schweizerbart Science: Stuttgart, pp 267–274.

Boenigk J, Arndt H. (2002). Bacterivory by heterotrophic flagellates: community structure and feeding strategies. *Antonie van Leeuwenhoek* 81: 465–480.

Boenigk J, Pflandl K, Garstebi T, Novarino C, Chatinotinas A. (2006). Evidence for geographic isolation and signs of endemism within a protistan morphospecies. *Appl Environ Microbiol* 72: 5159–5164.

Borcard D, Gillet F, Legendre P. (2011). *Numerical Ecology with R*. Springer: New York, Dordrecht, London, Heidelberg.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 6: 1621–1624.

Caron DA, Countway PD. (2009). Hypotheses on the role of the protistan rare biosphere in a changing world. *Aquat Microb Ecol* 57: 227–238.

Coleman DC, Whitman WB. (2004). Linking species richness, biodiversity and ecosystem function in soil systems. *Pedobiologia* 49: 479–497.

Colwell RK, Chao A, Gotelli NJ, Lin S-Y, Mao CX, Chazdon RL, Coleman DC, Whitman WB. (2004). Linking species richness, biodiversity and ecosystem function in soil systems. *Pedobiologia* 49: 479–497.

de Candolle AP. (1820). *Géographie botanique*. In: Cuvier et al., *Histoire naturelle des animaux sans vertèbres*, 13 vols, edns, Paris, pp 359–422.

Degnan PH, Ochman H. (2012). Illumina-based analysis of microbial community diversity. *ISME J* 6: 183–194.

Del Campo J, Massana R. (2011). Emerging diversity within chrysophytes, chaoflagellates and bicosoecids based on molecular surveys. *Protist* 162: 435–448.

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science* 348: 1261605.

De Wit R, Bouvier T. (2006). ‘Everything is everywhere, but the environment selects’: what did Baas Becking and Beijerinck really say? *Environ Microbiol* 8: 755–758.

Dokulil M, Hamm A, Kohl J-G. (2001). *Ökologie und Schutz von Seen*. Facultas-Universitäts-Verlag: Wien.

Edgar RC. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics* 26: 2460–2461.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27: 2194–2200.

Ehrenberg CG. (1849). Über das mächtigste bis jetzt bekannt gewordene (angeblich 500 füß mächtige) Lager von mikroskopischen reinen kieselgeschichtlichen Süsswasser-Formen am Wasserfall-Flusse im Oregon. *Berl. Bek. Verh. Königl.-Preuss. Akad. Wiss. Berlin*: 76–87.

Ehrenberg CG. (1850). On insular deposits on the River Chutes in Oregon. *Am J Sci* 9: 140.

Fenchel T. (1994). Microbial ecology on land and sea. *Philos Trans R Soc Ser B* 343: 51–56.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. (1998). Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* 281: 237–240.

Finlay BJ, Esteban GF. (1998). Freshwater protozoa: biodiversity and ecological function. *Biodivers Conserv* 7: 1163–1186.

Foissner W. (1991). Diversity and ecology of soil flagellates. In: Patterson DJ, Larsen J (ed), *The Biology of Free-living Heterotrophic Flagellates*, Vol Systematic Association Special Volume No 45. Clarendon Press and The Systematics Association: Oxford, pp 93–112.

Foissner W. (1999). Protist diversity: estimates of the near-imponderable. *Protist* 150: 363–368.

Foissner W. (2004). Ubiquity and cosmopolitism of protists questioned. *SILnews* 43: 6–7.

Foissner W. (2006). Biogeography and dispersal of micro-organisms: a review emphasizing protists. *Acta Protozool* 45: 111–136.

Geisen S, Tveit AT, Clark IM, Richter A, Svenning MM, Bonkowski M et al. (2015). Metatranscriptomic census of active protists in soils. *ISME J* 9: 2178–2190.

Glücksmann E, Bell T, Griffiths RI, Bass D. (2010). Closely related protist strains have different grazing impacts on natural bacterial communities. *Environ Microbiol* 12: 3105–3113.

Graham LE, Graham JM, Wilcox LW. (2009). *Algae*, 2nd edn. Pearson: San Francisco.

Ingemanison DE, Wallace WJ. (1995). *Oceanography – An Introduction*, 5th edn. International Thomson Publishing: London, p 272.

Katz LA, McManus GB, Snoeyenbos-West OLO, Griffin A, Pirog K, Costas B et al. (2005). Reframing the ‘Everything is everywhere’ debate: evidence for high gene flow and diversity in ciliate morphospecies. *Aquat Microb Ecol* 41: 55–65.

Lane D. (1991). *Nucleic Acid Techniques in Bacterial Systematics*, 1st edn. Wiley: Chichester, New York.

Legendre P, Gallagher ED. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia* 129: 271–280.

Lecroq B, Lejzerowicz F, Bachar D, Christen R, Esling P, Foissner W. (2010). *Protist diversity: estimates of the near-imponderable*. *Protist* 161: 435–448.

Logares R, Audic S, Bass D, Bittner L, Boutte C, Christen R et al. (2011). Ultra-deep sequencing of foraminiferal microbarcodes unveils hidden richness of early monothalamous lineages in deep-sea sediments. *Proc Natl Acad Sci* 108: 13177–13182.

Livermore JA, Jones SE. (2015). Local–global overlap in diversity informs mechanisms of bacterial biogeography. *ISME J* 9: 2413–2422.

Logares R, Audic S, Bass D, Bittner L, Boutte C, Christen R et al. (2014). Patterns of rare and abundant marine microbial eukaryotes. *Curr Biol* 24: 318–321.

Lopéz-García P, Rodríguez-Valera F, Pedrós-Alió C, Moreira D. (2001). Unexpected diversity of small
eukaryotes in deep-sea Antarctic plankton. Nature 409: 603–607.

Mahé F, Mayor J, Bunge J, Chi J, Siemensmeyer T, Stoeck T et al. (2014). Comparing high-throughput platforms for sequencing the V4 region of SSU-rDNA in environmental microbial eukaryotic diversity surveys. J Eukaryot Microbiol 62: 338–345.

Mann HB, Whitney DR. (1947). On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18: 50–60.

McMurdie PJ, Holmes S. (2014). Waste not, want not: why ratefying microbiome data is inadmissible. PLOS Comput Biol 10: 1–12.

Medinger R, Nolte V, Panday RV, Jost S, Ottenwälder B, Schlötterer C et al. (2010). Diversity in a hidden world: potential and limitation of next generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. Mol Ecol 19: 32–40.

Medlin L, Elwood HJ, Stickel S, Sogin ML. (1998). The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. Gene 71: 491–499.

Moreira D, López-García P. (2002). The molecular ecology of microbial eukaryotes unveils a hidden world. Trends Microbiol 10: 31–38.

Nolte V, Pandey RV, Jost S, Medinger R, Ottenwälder B, Boenigk J et al. (2010). Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. Mol Ecol 19: 2908–2915.

Novarino G, Warren A, Butler H, Lambourne G, Boxshall A, Bateman J et al. (1997). Protistan communities in aquifers: a review. FEMS Microbiol Ecol 20: 261–275.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O’Hara RB et al. (2013). Vegan: Community Ecology Package. R package version 2.0-10. Available from http://CRAN.R-project.org/package = vegan.

O’Malley MA. (2007). The nineteenth century roots of ‘everything is everywhere’. Nat Rev Microbiol 5: 647–651.

Pandey RV, Nolte V, Schlötterer C. (2010). CANGS: a user-friendly utility for processing and analyzing 454 GS-FLX data in biodiversity studies. BMC Res Notes 3: 3.

Ricci V. (2005). Fitting distributions with R. Available from http://cran.r-project.org/doc/contrib/Ricci-distri buttons-en.pdf.

Risse-Buhl U, Herrmann M, Lange P, Akob DM, Pizani N, Schonborn W et al. (2013). Phagotrophic protist diversity in the groundwater of a Karstified Aquifer—morphological and molecular analysis. J Eukaryot Microbiol 60: 467–479.

Stoeck T, Behnke A, Christen R, Amaral-Zettler L, Rodriguez-Mora MJ, Chistoserdov A et al. (2009). Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. BMC Biol 7: 72.

Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner H-W et al. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol 19: 21–31.

Tedersoo L, Bahram M, Põlme S, Kõljalg U, Yorou NS, Wijesundera R et al. (2014). Global diversity and geography of soil fungi. Science 346: 1256688.

Tikhonenkov DV, Mylnikov AP, Gong YCh, Feng WS, Mazei Y. (2012). Heterotrophic flagellates from freshwater and soil habitats in subtropical China (Wuhan Area, Hubei Province). Acta Protozool 51: 65–79.

Triadó-Margarit X, Casamayor EO. (2012). Genetic diversity of planktonic eukaryotes in high mountain lakes (central Pyrenees, Spain). Environ Microbiol 14: 2445–2456.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/