DR. TOMAS ROSLIN (Orcid ID : 0000-0002-2957-4791)

DR. NIELS MARTIN SCHMIDT (Orcid ID : 0000-0002-4166-6218)

PROF. OTSO OVASKAINEN (Orcid ID : 0000-0001-9750-4421)

Article type : Resource Article

SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes

Yinqiu Ji[1,*], Tea Huotari[2,*], Tomas Roslin[2,3], Niels Martin Schmidt[4,5], Jiaxin Wang[1], Douglas W. Yu[1,6,7,γ], Otso Ovaskainen[8,9,γ]

[1] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, 650223, China

[2] Spatial Foodweb Ecology Group, Department of Agricultural Sciences, University of Helsinki, P.O. Box 27, FI-00014 Helsinki, Finland

[3] Spatial Foodweb Ecology Group, Department of Ecology, Swedish Univ. of Agricultural Sciences, Box 7044, SE-750 07 Uppsala, Sweden

[4] Arctic Research Centre, Aarhus University, DK-8000, Aarhus, Denmark

[5] Department of Bioscience, Aarhus University, DK-4000, Roskilde, Denmark

[6] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk, NR47TJ UK

[7] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming Yunnan, 650223 China

[8] Organismal and Evolutionary Biology Research Programme, University of Helsinki, P.O. Box 65, FI-00014, Helsinki, Finland

[9] Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, N-7491, Trondheim, Norway

* Contributed equally to the paper

γ Corresponding authors

Statement of authorship: NMS has been involved in running the BioBasis sampling program for more than twenty years. TR, NMS, DWY, and OO conceived the study and its design. TH led the work in generating all the DNA samples and YJ led the work in assembling and annotating the mitogenomes for the mitochondrial genome reference database. TH led the work in generating the mock communities and bulk samples, with contributions from YJ and JW. YJ and DWY developed the molecular and bioinformatic methods. OO led the modelling of the data. TR and OO wrote the first draft of the manuscript, and all authors contributed substantially to its further improvement.

Data accessibility statement: Should the manuscript be accepted, the data supporting the results will be archived in an appropriate public repository (Dryad), and the data DOI will be included at the end of the article. The bioinformatic and R scripts and associated data tables will also be made available on github.com.

Running title: A metagenomic pipeline for community ecology

Key words: biomonitoring, community composition, metagenomics, mitogenomics, mitogenomes, COI internal standard, DNA barcoding, Arthropoda, Araneae, Insecta

Name and complete mailing address of the persons to whom correspondence should be sent:

Prof. Otso Ovaskainen (otso.ovaskainen@helsinki.fi, tel + 358-50-3092795)
Organismal and Evolutionary Biology Research Programme, University of Helsinki, P.O. Box 65, FI-00014, Helsinki, Finland
and
Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, N-7491, Trondheim, Norway

Prof. Douglas W. Yu (dougwyu@mac.com, tel +44-7510-308-272)
School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk, NR47TJ UK
and
State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, 650223, China
and

Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming Yunnan, 650223 China

Abstract

The accurate quantification of eukaryotic species abundances from bulk samples remains a key challenge for community ecology and environmental biomonitoring. We resolve this challenge by combining shotgun sequencing, mapping to reference DNA barcodes or to mitogenomes, and three correction factors: (1) a percent-coverage threshold to filter out false positives, (2) an internal-standard DNA spike-in to correct for stochasticity during sequencing, and (3) technical replicates to correct for stochasticity across sequencing runs. The SPIKEPIPE pipeline achieves a strikingly high accuracy of intraspecific abundance estimates (in terms of DNA mass) from samples of known composition (mapping to barcodes R^2=0.93, mitogenomes R^2=0.95) and a high repeatability across environmental-sample replicates (barcodes R^2=0.94, mitogenomes R^2=0.93). As proof of concept, we sequence arthropod samples from the High Arctic, systematically collected over 17 years, detecting changes in species richness, species-specific abundances, and phenology. SPIKEPIPE provides cost-efficient and reliable quantification of eukaryotic communities.

1. Introduction

The key dimensions of ecological community structure are species composition and species abundances (Vellend 2010). To survey communities of animals, plants, and fungi, researchers have traditionally relied on morphological characters. However, a technological revolution is beginning to replace this labor-intensive approach with spectral, acoustic, and

molecular data (Bohan *et al.* 2017; Bush *et al.* 2017; Ovaskainen *et al.* 2018; Schulte to Bühne *et al.* 2018), offering replicable and efficient ways to estimate change in community structure over time, space, and environmental gradients.

We focus here on the use of DNA-sequence information to estimate eukaryotic species composition and abundance from mixed-species samples, such as bulk samples of invertebrates (Ji *et al.* 2013; Hering *et al.* 2018; Pawlowski *et al.* 2018), or water or air samples that have been filtered to capture environmental DNA (Bohmann *et al.* 2014; Abrego *et al.* 2018; Goldberg *et al.* 2018). In particular, we aim at quantifying within-species, spatiotemporal variation in a set of known target species, as measured by variation in the occurrence and mass of marker DNA sequences. This is the type of information typically sought in applied biomonitoring.

The two main approaches to identify eukaryotic species from mixed-species samples are known as 'metabarcoding' and 'mitogenomics' (Tang *et al.* 2015; Crampton-Platt *et al.* 2016; Bush *et al.* 2017; Bista *et al.* 2017). Metabarcoding uses PCR to amplify short, taxonomically informative 'DNA barcode' sequences from mixed-species samples. These amplicons are then sequenced, and the reads are assigned to taxonomies by matching to barcode reference databases. Metabarcoding is associated with low per-sample cost, because samples are individually tagged during PCR so that multiple samples can be pooled before being prepared for sequencing ('library prep') and because even a low sequencing depth of a few thousand reads is sufficient for characterizing the dominant species of a sample. This means that many samples can be included in the same sequencing run. Metabarcoding also

has two other important advantages: access to well-populated reference databases, such as BOLD (Ratnasingham & Hebert 2007) and Midori (Machida *et al.* 2017), and the fact that PCR amplification allows the detection of species present only as low-concentration environmental DNA (eDNA), such as trace animal tissue in water. However, metabarcoding has two important limitations: susceptibility to sample contamination, due to PCR amplification of stray DNA (precisely what makes metabarcoding useful for eDNA in the first place), and loss of quantitative information, due to primer and polymerase biases (Yu *et al.* 2012; Piñol *et al.* 2015, 2018; Nichols *et al.* 2018; Deagle *et al.* 2018a; Lamb *et al.* 2019).

By contrast, mitogenomics is a variant of metagenomics and is based on the shotgun-sequencing of genomic DNA from a bulk sample, followed by mapping the reads onto a set of reference mitochondrial genomes, each of which serves as a 'super barcode' for a species (Crampton-Platt *et al.* 2016). Mitogenomics is more expensive than metabarcoding, because each sample must be individually library prepped, because samples must be sequenced more deeply than for metabarcoding, and because compiling a mitogenome reference database imposes additional costs for specimen acquisition, sequencing, and assembly. Mitogenomics is also unsuitable for eDNA because library prep requires sample DNA to be of high quality and quantity, and because mapping is only efficient if the sample consists mostly of the target organisms (but see Wilcox *et al.* 2018). However, mitogenomics has two key advantages over metabarcoding, due to not using PCR: mitogenomics is robust to sample contamination, because stray DNA is detected at low levels, if at all, and can thus be ignored. Also, mitogenomics preserves quantitative information, as the proportion of reads that map to a species' mitogenome correlates with that species' relative biomass in the

sample (Zhou *et al.* 2013; Gómez-Rodríguez *et al.* 2015; Tang *et al.* 2015; Bista *et al.* 2017).

Thus, mitogenomics promises the ability to measure *both* of a community's key dimensions, species composition and abundances. However, to date, reported correlations between read and abundance frequencies in samples of known composition have been noisy, and the shape of the relationship has varied idiosyncratically. For instance, analysing mixed-species communities, Zhou *et al.*'s (2013) pioneering study reported an R^2 of 63% for a curvilinear regression, Tang *et al.* (2015) reported a linear-regression R^2 of 25%, Gómez-Rodríguez *et al.* (2015) reported an R^2 of 64% for a linear log-log regression, and Bista *et al.* (2017) returned R^2 values between 45% and 87%, using logistic or linear models to fit different species.

Here we describe a step-change improvement in the mitogenomic pipeline, which performs well even when the mapping targets consist only of short DNA barcode sequences. To achieve accurate estimates of intraspecific variation in abundance, the pipeline employs a set of filters and internal standards: (1) a percent-coverage threshold to filter out false-positive mappings, (2) an internal-standard DNA spike-in to correct for sequencing-depth stochasticity across samples within a sequencing run, and (3) technical replicates to allow correction for sequencing-depth stochasticity across sequencing runs. We illustrate the pipeline by applying it to a time series (1997–2013) of pitfall-trap samples from the high Arctic, the Zackenberg Valley in Northeast Greenland (Schmidt *et al.* 2016a; Christensen *et al.* 2017).

2. Materials and methods

SPIKEPIPE consists of five steps, which span from the wet lab to bioinformatics to statistical analysis (Box 1), and which are co-designed to maximize accuracy in species detection and to preserve abundance information. Step 1 consists of compiling a barcode or mitogenome reference database to be used as the mapping target. Step 2 consists of constructing, standardizing, and sequencing a set of mock communities to be used for calibration. Step 3 consists of standardizing and sequencing environmental samples (here, the Zackenberg samples). Step 4 consists of mapping the reads from the mock communities and environmental samples against the reference database and using these mappings to compute predictor values for species occurrence and abundance. Step 5 consists of using the mock data to fit statistical models that predict species occurrences and abundances, and applying these models to predictor values computed from the environmental data. We use DNA mass (ng of mitochondrial DNA of a species in the extracted sample) as our abundance measure, which is a proxy for species biomass. Below, we first introduce the Zackenberg case study, then detail Steps 1–5, and finally evaluate and compare the performance when using DNA barcodes versus mitogenomes to estimate community structure, both for the mock and the real environmental samples.

Zackenberg Valley case study

The environmental dataset comes from a time series of yellow-pitfall-trap samples collected as part of the BioBasis monitoring program at the Zackenberg Research Station (Schmidt *et al.* 2016a; Christensen *et al.* 2017), located in the High-Arctic zone of northeast Greenland

(74°28’ N; 20°34’ W). Spiders and insects have been collected weekly throughout the summer from 1996 onwards from various tundra habitats. These samples have over the years been sorted to higher taxonomic rank (mostly family level), analyzed at this low taxonomic resolution (e.g. Høye *et al.* 2007, 2013, 2014; Reneerkens *et al.* 2016; Schmidt *et al.* 2016b, 2017), and warehoused as taxonomically sorted sub-samples in ethanol at room temperature in Museum of Natural History, Aarhus, Denmark (Supplement Text S3). The BioBasis protocols for sample collection, sorting, and storage have introduced cross-sample contamination (Supplement Text S3), making the samples unsuitable for metabarcoding. We use the subset of samples collected weekly in three yellow pitfall traps in a mesic heath habitat, from 1997 to 2013 (Supplement Text S3), during which time summer has doubled in length (see Kankaanpää *et al.* 2018).

Step 1. Compiling the DNA-barcode and mitogenome reference databases

For the DNA-barcode database, we use the 406 insect and spider BIN sequences (Barcode Identification Numbers) compiled by Wirta *et al.* (2016) at Zackenberg (lengths range from 453 to 654 bp, mean 613). For the mitogenome database, we assembled 308 mitogenomes (lengths range from 5768 to 18625, mean 15720). The protocol for assembling mitogenomes is standard (Gómez-Rodríguez *et al.* 2015; Tang *et al.* 2015; Crampton-Platt *et al.* 2016; Bista *et al.* 2017), so we provide here a summary, with details in Supplement Text S1. Each voucher specimen was individually library-prepped and shotgun-sequenced to prevent chimeric assemblies and to preserve the option of using the nuclear-DNA reads in the future (cf. Sarmashghi *et al.* 2018). We successfully assembled 283 mitogenomes from these data, with a further 25 mitogenomes assembled out of a mixture of these data and the mixed-

species pitfall-trap samples' data. Finally, we used COI from each mitogenome to confirm species identities, matching them to the BIN sequences of Wirta *et al.* (2016). Of the resulting 308 mitogenomes, 282 are in one contig, and the remaining 26 are in two or three contigs. 273 mitogenomes have all 13 protein-coding genes, and the remaining 35 have one or more missing or incomplete protein-coding genes (details in Supplement Table S1). 12S, 16S, and D-loop were seldom assembled, so these were omitted from all mitogenomes.

Step 2. Constructing, standardizing, and sequencing mock communities

To calibrate SPIKEPIPE and estimate its accuracy, we constructed two kinds of mock communities using known input-DNA amounts of Zackenberg species: 'mock-even' and 'mock-gradient' communities (detailed in Supplement Table S2). We created six mock-even communities. For each, we used 20 species (19 Diptera, one spider), with equal amounts of DNA from each of the 20 species. This taxonomic composition reflects realistic sample compositions in the High Arctic, which is dominated by Diptera (Wirta et al. 2016). To mimic variation in sample absolute biomasses, two mock-evens used 50 ng DNA per species, two used 100 ng, and two used 200 ng (each creating a technical-replicate pair). We also created three mock-gradient communities. For each, we used 19 input species (all Diptera), with the first species represented by 20 ng of DNA, increasing geometrically by a factor of 1.3 until the most abundant species, at 2698.9 ng, a 135-fold range. The mock-gradient samples mimic the common situation where a sample contains both abundant and rare species (here, by DNA mass). We further included two negative controls without any species. The pooled set of mock communities allowed us to address the following questions: (1) *replicability*: do technical replicates return the same results?; (2) *within-species*

quantification: for a given species, does the number of mapped reads correlate with the amount of input DNA?; (3) *across-species quantification*: do different species with the same input DNA return the same number of mapped reads?; and (4) *sensitivity*: can rare species be detected?

Internal-standard DNA – The fraction of the reads in a sample that maps to a species is an estimate of that species' relative biomass in the sample, but to estimate that species' absolute abundance (and allow comparisons across samples), we should correct for stochasticity in sequencing depth across samples. To do so, we used a fixed amount of an internal-standard DNA to 'spike' a 1.2 ml aliquot of lysis buffer from each of the mock samples (Box 1), so samples with more reads of the internal standard can be down-weighted (see Smets *et al.* 2016; Deagle *et al.* 2018b; Tkacz *et al.* 2018 for PCR-based protocols). To make the internal standard, we PCR-amplified and pooled fixed amounts of COI-barcode-amplicon DNA (658 bp) from three insect species collected in China and for which there are no confamilials in Greenland: 0.2 ng of *Bombyx mori* (Lepidoptera: Bombycidae), 0.4 ng of an unnamed beetle species (Coleoptera: Mordellidae), and 0.8 ng of another unnamed beetle species (Coleoptera: Elateridae) (For the PCR protocol, see Supplement Text S2.2). We used three species to check for error and degradation; the mapped reads from the three species should be found in the input ratio of 1:2:4. If, for instance, one of the species is not at the correct ratio with the other two, we can omit that species' reads. After spiking each mock sample, we extracted and purified the DNA and sent the mock samples for individual library prep and sequencing at the Earlham Institute (Supplement Text S2 for details). We sequenced the 6 mock-evens twice because one of the 50-ng mocks failed the first time,

resulting in 11 total mock-even and 3 mock-gradient datasets, as well as the 2 negative controls.

Step 3. Constructing, standardizing, and sequencing environmental samples

In 2016 and 2017, we non-destructively extracted DNA from the warehoused taxonomic subsets of 493 trap-week samples collected from 1997 to 2013 (except for 2010, the samples of which had been lost in transport from Greenland) (Supplement Text S4.1). We then pooled lysis-buffer solutions to reconstruct the 493 complete trap-week samples, aliquoted 1.2 ml from each sample, spiked each aliquot with 0.2 ng, 0.4 ng, and 0.8 ng of the three internal-standard species (from Step 2), extracted and purified the DNA, and sent them for individual library prep and sequencing at the Earlham Institute. The samples were spread over three sequencing runs, so we included 4 or 5 technical replicates per year of samples from the previous two runs in the final run, to allow correction for stochasticity in sequencing depth across runs (Supplement Text S4.2).

In a variant protocol, for years 1997-1999 and 2011-2013, the first samples that we sequenced, we used 10 ng, 20 ng, and 40 ng of the three internal-standard species. We subsequently determined that the internal-standard DNA represented a too-large proportion of the resulting reads (mean 10.9%), so we library-prepped and sequenced the remaining DNA from the aliquots a second time to increase data. Altogether, we successfully sequenced 728 samples: 14 mocks, 712 trap-week samples (original and technical-repeat aliquots), and two negative controls (which returned only the internal-standard species) (Supplement Table S4.2.1).

Step 4. Mapping reads and computing predictors

All sequencing outputs (mock communities, negative controls, and the Zackenberg samples) were processed as follows (details in Supplement Text S5). We removed Illumina adapter sequences using *TrimGalore* 0.4.5 (Martin 2011) and used *minimap2* 2.10 in short-read mode (Li 2018) to map the paired-end reads against either the DNA-barcode or the mitogenome reference database, plus the three internal-standard sequences. We used *samtools* 1.5 (Li *et al.* 2009) to exclude reads that mapped as secondary or supplementary alignments and to include only paired-end reads mapped at quality ≥48, in the correct orientation, and at approximately the correct distance apart ('proper pairs').

PC, FSL, RUN, SPECIES – If a species is truly present in a sample, we expect reads from that sample *to map along the length* of that species' DNA-barcode or mitogenome, not just to one segment, resulting in a higher percentage coverage (henceforth *PC*) (Supplement Figure S5.1). We used *bedtools* 2.27.1 (Quinlan & Hall 2010) to calculate PC as the fraction of positions covered by one or more mapped reads. We expected the fraction of reads from a sample that map to a species' barcode or mitogenome to be correlated with that species' relative abundance. However, the absolute number of mapped reads in a sample is determined by that sample's dataset size, which is affected by numerous random factors from DNA extraction to sequencing. We used an internal standard with the aim of correcting for these random variations and recovering each sample's original biomass, allowing calculation of each species' absolute abundance per sample (Box 1). To do so, we added the same amount of internal-standard DNA to the fixed aliquot of each sample before DNA extraction (see *Internal-standard DNA*), and kept track of the fraction of total lysis buffer

that the aliquot represents. We defined the quantity *FSL* for each species in each sample, defined as $\log(F/SL)$, where F is the number of reads mapped to a focal species (unit: sequence count), S is the "spike" (unit: sequence count/DNA mass), and L is the fraction of lysis buffer represented by the aliquot (unitless). The spike S was first computed for each of the three internal standards as number of reads mapped to the internal standard divided by the input amount of DNA, and then averaged over the three internal standards. To test the importance of the internal standard, we also computed *FTL*, defined as $\log(F/TL)$, where F and L are the same as above, and T is the total number of sequences in a sample's dataset. As the mapping from *PC* and *FSL* to species presences and abundances might also vary among sequencing runs, we used the identity of the three sequencing runs (henceforth RUN), as an additional predictor, implemented as a factor. Finally, as sequencing success could be idiosyncratically species-specific, e.g. species varying in their mitochondrial-to-nuclear-DNA ratios or in how much tissue is released into solution during lysis, we considered the identity of the species (henceforth *SPECIES*) as an additional predictor, implemented as a factor. To summarize, the outcome of Step 4 is two datasets of values for *PC*, *FSL*, *RUN,* and *SPECIES*, computed after mapping reads from the pooled mock-community dataset against the DNA-barcode and the mitogenome reference databases.

Step 5. Predicting species occurrences and abundances

In this step, statistical models that predict species occurrences and abundances were calibrated with the mock-sample datasets and then applied to the environmental data. First, we used logistic regression to model species presence/absence as a function of *PC*, *FSL*, *RUN*, and *SPECIES*, of which we expected *PC* to be the most important predictor. Second, we

used a linear regression with the same predictors to model variation in abundance conditional on presence (the log-transformed amount of DNA included in the mock community), with the expectation of *FSL* being the most important predictor. Note that we used log-transforms in our models because the amounts of DNA in both the mock samples and the environmental samples can vary over orders of magnitude. In such a case, the researcher is likely to be interested in quantifying proportional differences among samples, e.g. “sample A has 100 times more and sample B has 10 times more DNA of the focal species than sample C has”. Moreover, we log-transformed both the response variable and the main predictor (FSL) to make them compatible with each other. Thus, the theoretical expectation for the slope value of *FSL* is 1.0, because that corresponds to the number of (SL-corrected) mapped reads being directly proportional to the amount of input DNA.

To examine the information gained from the addition of the spike, we replaced in the best model *FSL* by *FTL*. To examine how much accuracy is lost if the results are not calibrated per run, we dropped *RUN* from the model. To examine how well the abundance estimates are calibrated across species, we examined performance of a model variant where *SPECIES* was dropped. Further, to examine if species effects can be explained by their mitogenome lengths, we regressed the estimated effects of *SPECIES* from the abundance model on log-transformed mitogenome length, reasoning that longer mitogenomes should attract more reads. Finally, the models calibrated with mock community data were used to predict species occurrences and abundances in the environmental data (Box 1).

Evaluating SPIKEPIPE's replicability with re-sequenced samples

After performing steps 1–5 of SPIKEPIPE described above, we tested its replicability by comparing results between technical replicates. If a sample was sequenced more than twice (see Table S4.2.1), we considered all pairs of replicates, so that a sample sequenced in runs X, Y and Z contributed to the comparisons (X,Y), (X,Z) and (Y,Z). To evaluate the robustness of detecting species presence, we counted how often a species was inferred to be present or absent in both runs of comparison (agreement) or present in only one (disagreement). To evaluate the robustness of species abundance estimation, we used only cases where a species was classified as present in both runs of comparison. For each pair of runs, we used a linear model to estimate how well the abundance estimates obtained from one run explained variance in abundance estimates in the second run. We performed this analysis either controlling or not controlling for the effect of the specific run, to examine whether run-specific correction factors are needed.

Evaluating SPIKEPIPE's utility for making ecological inference

Finally, we tested SPIKEPIPE's power to detect ecological change, and how this power differs between mapping against reference DNA barcodes or reference mitogenomes. To this aim, we merged the results from the four runs by assuming that a species is present in a sample if it was inferred to be present in any of the runs containing that sample, and we used the average of run-corrected abundance estimates as the final estimate for abundance. We then normalized abundance within each species across all samples to mean zero and unit variance, so that the unit of the abundance is within-species deviation from the species' mean. We summed inferred presences over all species to estimate species richness per trap-

week, and we averaged normalized abundances over all species to estimate community-wide variation in abundance. To test whether these richness and abundance data convey signals of change in a warming Arctic, we used Poisson regression to model species richness per trap-week, and linear regression to model mean abundance conditional on presence. In both cases, the data points (n=493) are weekly samples from the three traps in the years 1997–2013. We used as the candidate predictors the continuous effect of the year (to capture a linear trend), the Julian date and its square (to capture phenological variation within year), the interaction between year and Julian date (to capture change in phenological timing), and the trap (to capture technical variation among the three traps). We performed variable selection in all models using the Akaike Information Criterion (AIC).

3. Results

3.1. Mock community analyses: species composition and abundances

Mapping to short barcodes and long mitogenomes were both highly successful in identifying species occurrences in the mock samples, even though these included the mock-gradients with their rare species: the barcode approach detected 97%, and the mitogenome approach detected 100% of the species that were actually present (Table 1). The false-positive rates were also very low even when using the lax criterion of letting just one mapped read infer species presence (Table 1). Due to this high separation, fitting logistic regression models was not meaningful. Instead, we defined conservative thresholds PC_0 for percentage cover (set to PC_0=0.5 for barcodes and PC_0=0.1 for mitogenomes), so that applying the threshold condition $PC > PC_0$ resulted in zero false positives and thus gave strong evidence for species

occurrence in a sample (Table 1, see also Tang *et al.* 2015). Sensitivity remained high when applying the criterion $PC > PC_0$, with the barcode approach detecting 95% and the mitogenomics approach 97% of the species that were actually present.

In terms of sensitivity, we found a taxonomic difference. Among the two classes represented in the mock communities (19 Diptera, one spider), the spider performed worse than the other taxa in our analyses: in the mock data, the spider was present in 11 samples and absent in 5 samples. For the barcode approach, the presence-absence model predicted that the species was present only once, resulting in 1 true positive, 5 true negatives, and 10 false negatives. For the mitogenomics approach, the number of true positives was 3, and thus the number of false negatives was 8. This implies that further tests with species outside Diptera will be needed to validate the generality of our results.

Turning next to species abundances, we focused on the species inferred to be present by $PC > PC_0$, attempting to predict how abundant each was. Model selection with AIC supported keeping *FSL*, *RUN*, and *SPECIES* in the final model when mapping against either barcodes or mitogenomes. With these models, R^2 equaled 0.93 for barcodes and 0.95 for mitogenomes (Table 2, Box 1), and the slope of *FSL* was 0.90 for both barcodes and mitogenomes, i.e. close to the theoretical expectation of 1.0. Models without the internal standard (i.e. using *FTL* rather than *FSL* as the predictor) performed poorly, as did models from which *RUN* or *SPECIES* was dropped (Table 2, Box 1). Variation among *SPECIES* effects in the final model was not explained by length of barcode (p=0.96) nor mitogenome (p=0.21). In total, these results imply that calibration between sequencing runs is necessary and that only *within*-

species variation in abundance can be quantified accurately (i.e. 'Species A is more abundant in this sample than in another'). To estimate *among*-species variation (i.e. 'Species A is more abundant than species B'), it will be necessary to run mock communities with all target species of interest to estimate species-specific calibration factors.

3.2. Zackenberg analyses: replicability

Data based on mitogenomes provided a somewhat higher success rates than did data based on barcodes, as judged by two criteria: a higher number of species presences inferred, and a lower proportion of cases where a species was inferred to be present in one run only (Table 3). But overall, the numbers were surprisingly similar (Table 3), given that mean barcode length was only 3.9% of mean mitogenome length. The comparisons also suggest that the error rate in identifying species presence is somewhat higher in the environmental data (Table 3) than with the mock communities (Table 1), as may be expected given extra complexities with environmental data. The consistency of the run-corrected abundance estimates was almost as high as with the mock data: the abundance estimate from one run explained 94% (for barcodes) or 93% (for mitogenomes) of the variation of the abundance estimate derived from the other run (Fig. 1AD). Species of different taxonomic affinity varied slightly in performance (see Supplement S4.3). The consistency among replicate samples was somewhat higher for Araneae than for Diptera, and somewhat lower for Hymenoptera and Lepidoptera than for Diptera, but these latter two orders were generally too rare to make robust conclusions.

3.3. Zackenberg analyses: ecological inference

Of the 371 species that were present in our DNA-barcode reference database and that were systematically sampled over the years, we observed across all samples 145 species at least once, and 72 species at least five times. Of the 308 species in our mitogenome reference database, we observed 148 species at least once, and 81 species at least five times. Even though our analyses with mitogenome reference database contained 17% fewer species, mapping to mitogenomes consistently revealed higher species richness (53 species observed on average per year) than did mapping to DNA barcodes (46 species observed on average per year; Fig. 2).

We then contrasted the effects of using barcodes versus mitogenomes on estimates of community change, including only species that were observed at least five times. In our models of species richness *per trap-week*, we observed a peak in the middle of the summer and a decrease from 1997 to 2013 (Fig. 1BE), regardless of whether we used barcodes or mitogenomes. With mitogenomes (but not with barcodes), a statistical interaction between year and Julian season was retained, suggesting that peak species richness occurred earlier in 2013 than in 1997 (Fig. 1E). Both the barcode and the mitogenome datasets recorded greater total species abundance in 2013 over 1997, but only the mitogenome approach detected that community abundance decreased over the summer (Fig. 1CF).

4. Discussion

The harnessing of high-throughput DNA sequencing to infer species composition and species abundances is a huge opportunity for community ecology and environmental monitoring (Tang *et al.* 2015; Bush *et al.* 2017; Bista *et al.* 2017; Porter & Hajibabaei 2018). We report a step-change improvement in the accuracy and precision of the mitogenomic approach, achieved by combining shotgun sequencing with a percent-coverage threshold (PC_0), an internal-standard DNA spike-in (*FSL*), and correction factors among sequencing runs (*RUN*; Box 1). The highest accuracy and precision are achieved with longer mitogenome targets but remain surprisingly high with short DNA-barcode targets (Table 1, 3; Box 1, Fig. 1), and both approaches recover signals of structural change from real samples of arthropods at a high-Arctic site with marked environmental change (Figs. 1, 2).

4.1. High accuracy and precision

Our pipeline accurately inferred the presence and absence of species in mock samples. By selecting a conservative threshold for percentage cover (Table 1), we could completely eliminate false positives, while keeping false-negative rates only at 3% and 5% when mapping to mitogenomes and DNA barcodes, respectively. However, differences in the performance of Diptera and Araneae suggest that further tests will be needed to validate the generality of our results across a wider range of taxa. We could also accurately estimate within-species variation in abundance, achieving R^2 values of 0.95 for mitogenomes and 0.93 for COI barcodes, and reaching almost direct proportionality between the number of mapped reads and species abundance (Fig. 1). These results are a major improvement over

many previous studies (including our own), which have reported positive but noisy and idiosyncratic correlations between read number and biomass (Tang *et al.* 2015, Zhou *et al.* 2013; Gómez-Rodríguez *et al.* 2015; Bista *et al.* 2017). Overall, they seem to deliver on the key promise of mitogenomic approaches, adding quantitative information beyond that achieved by metabarcoding approaches (see Introduction).

We used as a unit of abundance DNA mass – i.e. nanograms of mitochondrial DNA of each species in a compound sample of DNA. Importantly, this is an absolute rather than a relative metric, allowing one to assess variation in the occurrence and concentration of DNA of the target species among samples, and thus to estimate within-species population changes. By measuring such changes for multiple species, one can detect quantitative changes in community structure. If desired, the unit of DNA mass can be converted into an estimate of the more traditional metric of individual counts per sample, given the availability of an estimate of mitochondrial DNA mass per individual per species.

Like any survey method, the method presented here will likely miss some species that are actually present in an environmental sample (Table 3), especially those present at low abundance. Thus, exhaustive sampling will require several sequencing runs, or at least a very high sequencing depth in a single run. When multiple sequencing runs are used, our results suggest that variation between runs can be almost completely accounted for, if including some samples in all runs to enable calibration. However, our tests are not exhaustive, and that this conclusion should thus be taken with caution. We also note that our mock samples were designed to reflect the High-Arctic fauna, which is dominated by the order, Diptera,

and we have shown that it is possible to differentiate congener species in this order, which is an important requirement for understanding changes in this community. However, we caution that given the enormous diversity of the Arthropoda, our tests can only be considered partial, and we robustness should continue to be tested in other studies.

4.2 Mitogenomes versus DNA barcodes

We consistently detected more species when mapping to mitogenomes than to barcodes – even though we had 406 barcodes but only 308 mitogenomes (Table 1, Fig. 2). This difference was caused by the higher PC threshold needed to avoid false positives for barcodes (PC_0 = 0.5) than for mitogenomes (PC_0 = 0.1). Perhaps as a result, mitogenome datasets also appear to be more powerful for detecting change in community structure (Fig. 1EF).

However, the higher statistical power of a mitogenome reference dataset trades off against the higher cost of compiling it, especially if suitable barcode sequences are already available (Ratnasingham & Hebert 2007; Machida *et al.* 2018; Nilsson *et al.* 2018). A mitogenome reference dataset is likely justified if the aim is to repeatedly monitor a fixed species list, in which case the initial cost of assembling a mitogenome dataset can be amortized. Conversely, in species-rich areas, DNA barcodes might be the only feasible option, in which case our results suggest that it will be helpful to increase sequencing depth per sample, and possibly also the number of samples per location. For example, the new Illumina NovaSeq delivers 1.5–5 times the output of the Illumina 2500 that we used (Bleidorn 2017). Our per-sample sequencing cost on the Illumina 2500 ranged from 130€ to 220€ per sample, plus the

LITE library cost of 5.6€ per sample. A high sequencing depth was required because in our samples, only ~0.2-0.4% of the reads mapped to our 308 mitogenomes, the remainder representing the nuclear genomes or microbes, not counting the internal-standards. Baits could potentially help (Liu *et al.* 2015, Wilcox *et al.* 2018), but baits do introduce bias, and the enrichment factor will vary across samples, obscuring abundance information. A final consideration is that a barcode reference dataset is currently more likely to allow detection of unexpected species, since barcode datasets can more easily include species from outside the study area. However, it could eventually become routine to carry out high-volume *de novo* mitogenome assembly from bulk samples themselves (Crampton-Platt *et al.* 2015), allowing species discovery and detection from the same samples. A related strategy is targeted assembly of barcodes from shotgun-sequenced samples (Greenfield et al. 2019).

4.3. Changes in community structure correlated with Arctic warming

SPIKEPIPE revealed major changes in mean phenology and abundance in the Zackenberg arthropod community. While we will report detailed analyses elsewhere, the patterns identified in Fig. 1BCEF expand upon two morphology-based studies in Zackenberg. Loboda *et al.* (2018) individually identified 18,385 Muscidae flies into 16 species and found a decline in species diversity between 1996 and 2014 – but no statistically significant change in species abundances in the mesic heath habitat (the same habitat as our environmental samples). Bowden *et al.* (2018) individually identified 28,566 spiders into nine species and found a decline in some species but no change in others. Such differential responses across species are altering the structure of Arctic communities (Høye *et al.* 2014; Kankaanpää *et al.* 2018, Koltz *et al.* 2018), and the changes are percolating to functional associations, e.g.

between plants and their pollinators (Schmidt *et al.* 2016b; Cirtwill *et al.* 2018). We clearly need species-resolution time series from full communities in order to understand the effects of climate change on biodiversity. An expansion of the morphological approach is infeasible, but our improvements in quantification, coupled with recent and imminent gains in cost-efficiency, now make the mitogenomic approach an attractive option for use in community ecology and in applied biomonitoring.

Acknowledgments

Arthropod samples were provided by the Greenland Ecosystem Monitoring Programme. The authors gratefully acknowledge CSC – IT Center for Science Finland and the HPC team at UEA UK, for computational resources. We thank the Danish National High-Throughput DNA Sequencing Centre and the Earlham Institute for sequencing. We thank Jaana Kekkonen and Isabella Palorinne for help during the laboratory work. We received funding from the Academy of Finland (grants 276909 and 285803 to TR, grants 284601 and 309571 to OO), the Jane and Aatos Erkko Foundation Grant (OO and TR), the Research Council of Norway through its Centres of Excellence Funding Scheme (223257) to OO via Centre for Biodiversity Dynamics. We are indebted to the Danish Environmental Protection Agency for funding BioBasis Zackenberg over the years, and to the many field and lab assistants over the years. D.W. Yu and Y.Q. Ji were supported by the National Natural Science Foundation of China (41661144002, 31670536, 31400470, 31500305), the Key Research Program of Frontier Sciences, CAS (QYZDY-SSW-SMC024), the Bureau of International Cooperation (GJHZ1754), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA20050202, XDB31000000), the Ministry of Science and Technology of China (2012FY110800), the State

Key Laboratory of Genetic Resources and Evolution (GREKF18-04) at the Kunming Institute of Zoology, the University of East Anglia, and the University of Chinese Academy of Sciences. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supporting Information

Supporting Information includes additional information on constructing the mitochondrial genome reference database (Text S1), constructing, standardizing, and sequencing mock communities (Text S2), the environmental samples (Text S3), processing of bulk samples from the Zackenberg collection (Text S4), mapping reads against references (Text S5), and instructions on how to download the example data and run Step 4 (bioinformatics) of the pipeline (Text S6), and how to run Step 5 (statistics) of the pipeline (Text S7).

References

Abrego, N., Norros, V., Halme, P., Somervuo, P., Ali-Kovero, H. & Ovaskainen, O. (2018) Give me a sample of air and I will tell which species are found from your region: Molecular identification of fungi from airborne spore samples. *Mol. Ecol. Resour.*, 18, 511–524.

Bista, I., Carvalho, G.R., Tang, M., Walsh, K, Zhou, X., Hajibabaei, M. *et al.* (2017) Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Mol. Ecol. Res.,* 18, 1020–1034.

Bleidorn C. (2017) *Phylogenomics*. Springer, Cham. pp. 43–60.

Bohan, D.A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J. & Woodward, G. (2017) Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends Ecol. Evol.,* 32, 477–487.

Bohmann, K., Evans, A., Gilbert, M.T.P., Carvalho, G.R., Creer, S., Knapp, M. *et al.* (2014) Environmental DNA for wildlife biology and biodiversity monitoring. *Trends Ecol. Evol.,* 29, 358–367.

Bowden, J.L, Hansen, O.L.P., Olsen, K., Schmidt, N.M. & Høye, T.T. (2018) Drivers of inter-annual variation and long-term change in High-Arctic spider species abundances. *Polar Biol.,* 41, 1635–1649.

Bush, A., Sollmann, R., Wilting, A., Bohmann, K., Cole, B., Balzter, H., *et al.* (2017) Connecting Earth observation to high-throughput biodiversity data. *Nature Ecol. Evol.*, 1*,* 0176.

Christensen, T.R., Topp-Jørgensen, E., Sejr, M.K. & Schmidt, N.M. (eds.) (2017) Special Issue: Synthesis of the Greenland Ecosystem Monitoring Programme. *Ambio* 46*,* 1.

Cirtwill, A.R., Roslin, T., Rasmussen, C., Olesen, J.M., & Stouffer, D.B. (2018) Between–year changes in community composition shape species' roles in an Arctic plant–pollinator network. *Oikos,* 127, 1163–1176.

Crampton-Platt, A., Timmermans, M.J.T.N., Gimmel, M.L., Kutty, S.N., Cockerill, T.D., Khen, C.V. *et al*. (2015) Soup to tree: the phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. *Mol. Biol. Evol*., 32, 2302–2316.

Crampton-Platt, A., Yu, D.W., Zhou, X. & Vogler, A.P. (2016) Mitochondrial metagenomics: letting the genes out of the bottle. *GigaScience*, 5*,* 15.

Deagle, B.E., Thomas, A.C., McInnes, J.C., Clarke, L.J., Vesterinen, E.J., Clare, E.L. *et al.*

(2018a) Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data?. *Mol. Ecol.*, https://doi.org/10.1111/mec.14734

Deagle, B.E., Clarke, L.J., Kitchener, J.A., Polanowski, A.M. & Davidson, A.T. (2018b) Genetic monitoring of open ocean biodiversity: An evaluation of DNA metabarcoding for processing continuous plankton recorder samples. *Mol. Ecol. Resour.*, *18*, 391–406.

Goldberg, C.S., Strickler, K.M. & Fremier, A.K. (2018) Degradation and dispersion limit environmental DNA detection of rare amphibians in wetlands: Increasing efficacy of sampling designs. *Sci. Total Environ.,* 633, 695–703.

Gómez-Rodríguez, C., Crampton-Platt, A., Timmermans, M.J., Baselga, A. & Vogler, A.P. (2015) Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods Ecol. Evol.,* 6, 883–894.

Greenfield, P., N. Tran-Dinh, and D. Midgley. 2019. Kelpie: generating full-length "amplicons" from whole-metagenome datasets. PeerJ 6:e6174.

Hering, D., Borja, Á., Jones, J.I., Pont, D., Boets, P. , Bouchez, A. *et al.* (2018) Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Research,* 138, 192–205.

Høye, T.T., Post, E., Meltofte, H., Schmidt, N.M. & Forchhammer, M.C. 2007. Rapid advancement of spring in the High Arctic. *Curr. Biol.,* 17, R449–R451.

Høye, T.T., Post, E., Schmidt, N.M., Trøjelsgaard, K., & Forchhammer, M.C. (2013) Shorter flowering seasons and declining abundance of flower visitors in a warmer Arctic. *Nat. Clim. Change,* 3, 759–763.

Høye, T.T., Eskildsen, A., Hansen, R.R., Bowden, J.J., Schmidt, N.M. & Kissling, W.D. (2014)

Phenology of high-arctic butterflies and their floral resources: Species-specific responses to climate change. *Curr. Zool.,* 60*,* 243–251.

Ji, Y.Q., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., *et al*. (2013) Reliable, verifiable, and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett.,* 16, 1245–1257.

Kankaanpää, T., Abrego, N., Skov, K., Lund, M., Schmidt, N. M., & Roslin, T. (2018) Spatiotemporal snowmelt patterns within High Arctic landscape - with implications for flora and fauna. *Arct. Antarct. Alp. Res., 50*, e1415624.

Koltz, A.M., Schmidt, N.M. & Høye, T.T. (2018) Differential arthropod responses to warming are altering the structure of Arctic communities. *Royal Soc. Open Sci.,* 5*,* 171503.

Lamb, P., Hunter, E., Pinnegar, J., Creer, S., Davies, R., & Taylor, M. (2019) How quantitative is metabarcoding: A meta-analytical approach? *Mol. Ecol.* 28(2):420-430.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics,* 3*,* 321.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. *et al*. (2009) The Sequence alignment/map format and SAMtools*. Bioinformatics*, 25, 2078–9.

Liu, S., Wang, X., Xie, L., Tan, M., Li, Z., Su, X. *et al*. 2015. Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Mol. Ecol. Resour.*, 16*,* 470–479.

Loboda, S., Savage, J., Buddle, C.M., Schmidt, N.M. & Høye, T.T. (2018) Declining diversity and abundance of High Arctic fly assemblages over two decades of rapid climate warming. *Ecography*, 41, 265–277.

Machida, R.J., Leray, M., Ho, S.L. & Knowlton, N. (2017) Metazoan mitochondrial gene

sequence reference datasets for taxonomic assignment of environmental samples. *Scientific Data,* 4, 170027.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, 17, 10–12.

Nichols, R.V., Vollmers, C., Newsom, L.A., Wang, Y., Heintzman, P.D., Leighton, M. *et al*. (2018) Minimizing polymerase biases in metabarcoding. *Mol. Ecol. Resour.*, 18*,* 927–939.

Nilsson, R.H., Larsson, K.-H., Taylor, A.F.S., Bengtsson-Palme, J., Jeppesen, T.S., Schigel, D., *et al.* (2018) The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res.*, DOI: 10.1093/nar/gky1022.

Ovaskainen, O., Camargo, U., Somervuo, P. 2018. Animal Sound Identifier (ASI): software for automated identification of vocal animals. *Ecol. Lett.*, 21, 1244–1254.

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A. *et al*. (2018) The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Sci. Total Environ.,* 637–638*,*1–16.

Piñol, J., Mir, G., Gomez-Polo, P. & Agustí, N. (2015) Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Mol. Ecol. Res.*, *15*, 819–830.

Piñol, J., Senar, M. A., & Symondson, W. O. (2018) The choice of universal primers and the characteristics of the species mixture determines when DNA metabarcoding can be quantitative. *Mol.Ecol.,* https://doi.org/10.1111/mec.14776.

Porter, T.M. & Hajibabaei, M. (2018) Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Mol.Ecol*., 27, 313–338.

Quinlan, A.R. & Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26*,* 841–842.

Ratnasingham, S. & Hebert, P.D. (2007) BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Mol. Ecol. Resour.*, 7, 355–364.

Reneerkens, J., Schmidt, N.M., Gilg, O., Hansen, J., Hansen, L.H., Moreau, J. & Piersma, T. (2016) Effects of food abundance and early clutch predation on reproductive timing in a high Arctic shorebird exposed to advancements in arthropod abundance. *Ecol. Evol.*, *6*(20), 7375–7386.

Sarmashghi, S., Bohmann, K., Gilbert, M.T.P., Bafna, V. & Mirarab, S. (2018) Assembly-free and alignment-free sample identification using genome skims. bioRxiv doi: https://doi.org/10.1101/230409

Schmidt, N.M., Hansen, L.H., Hansen, J., Berg, T.B. & Meltofte, H. (eds.) (2016a) *BioBasis, Conceptual design and sampling procedures of the biological monitoring programme within Zackenberg Basic*. 19th ed. Aarhus University, DK.

Schmidt, N.M., Mosbacher, J.B., Nielsen, P.S., Rasmussen, C., Høye, T.T. & Roslin, T. (2016b) An ecological function in crisis? – the temporal overlap between plant flowering and pollinator function shrinks as the Arctic warms. *Ecography,* 39, 1250–1252.

Schmidt, N.M., Hardwick, B., Gilg, O., Høye, T.T., Krogh, P.H., Meltofte, H. *et al.* (2017) Interaction webs in arctic ecosystems: Determinants of arctic change? *Ambio*, 46, Supplement 1*,* 12–25.

Schulte to Bühne, H. & Pettorelli, N. (2018) Better together: Integrating and fusing multispectral and radar satellite imagery to inform biodiversity monitoring, ecological research and conservation science. *Methods Ecol. Evol.,* 9, 849–865.

Smets, W., Leff, J.W., Bradford, M.A., McCulley, R.L., Lebeer, S. & Fierer, N. (2016) A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil. Biol. Biochem.,* 96*,* 145–151.

Tang, M., Hardman, C.J., Ji, Y., Meng, G., Liu, S., Tan, M. *et al.* (2015) High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods Ecol. Evol.,* 6, 1034–1043.

Tkacz, A., Hortala, M. & Poole, P.S. (2018) Absolute quantitation of microbiota abundance in environmental samples. *Microbiome* 6*,* 110.

Vellend, M. (2010) Conceptual synthesis in community ecology. *Q. Rev. Biol*., 85, 183–206.

Wilcox, T.M., Zarn, K.E., Piggott, M.P., Young, M.K., McKelvey, K.S. & Schwartz, M.K. (2018) Capture enrichment of aquatic environmental DNA: A first proof of concept. *Mol. Ecol. Resour.*, 18, 1392–1401.

Wirta, H.K., Várkonyi, G., Rasmussen, C., Kaartinen, R., Schmidt, N.M., Hebert, P.D.N. *et al.* (2016) Establishing a community-wide DNA barcode library as a new tool for arctic research. *Mol. Ecol. Resour.*, *16*, 809–822.

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. *et al.* (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods Ecol. Evol.,* 3, 613–623.

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L. *et al.* (2013) Ultra-deep sequencing enables

high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience* 2*,* 4.

Box 1. An overview of the SPIKEPIPE metagenomic pipeline for quantifying eukaryotic species presences and abundances from mixed-species bulk samples. SPIKEPIPE maps shotgun-sequenced reads from the sample against whole or partial mitogenomes, including DNA barcode sequences (typically the 3' portion of mtCOI for animals). Step 1 consists of compiling the barcode or mitogenome reference database (for barcodes, reference sequences can be downloaded from global databases such as BOLD; Ratnasingham & Hebert 2007). Step 2 consists of constructing, standardizing, and sequencing a set of mock communities to calibrate SPIKEPIPE. Step 3 consists of standardizing and sequencing a set of environmental bulk samples. Step 4 consists of mapping the reads from the mock communities and environmental samples against the reference database and using these mapping data to compute predictor values of species occurrence and abundance: *PC*, *FSL*, *SPECIES*, *RUN*. Step 5 uses the predictor values computed for the mock communities, for which we know the true species occurrences and abundances, Y(mock), to fit statistical models that predict species occurrences and abundances. These calibrated models are then applied to predictor values computed from the environmental mapping data to estimate species occurrences and abundances in the environmental samples, Y(environmental). Finally, the estimated species occurrences and abundances are used for modelling community change (Figs. 1, 2).

	Barcode (+)	Barcode (-)	Mitogenome (+)	Mitogenome (-)
True cases	277	6219	277	4651
≥ 1 read	269 (97%)	31 (0.5%)	277 (100%)	138 (3%)
PC > PC_0	264 (95%)	0 (0%)	269 (97%)	0 (0%)

Table 1. True- and false-positive and negative observations of species occurrence in the mock-community experiment. The table shows the total numbers of species that were present (+) or absent (-), pooled over the 14 mock-communities (True cases). The next two rows report the number of inferred species presences based on (1) the lax criterion that ≥ 1 reads were mapped to a focal species and (2) the stringent criterion that mapping percent-coverage, PC, exceeded a threshold PC_0. Note that even with the lax criterion, only 0.5% (barcode targets) or 3% (mitogenome targets) of truly absent species were identified falsely as present. With the stringent criterion, there were zero false positives.

Model	Barcode		Mitogenome	
	R^2	ΔAIC	R^2	ΔAIC
Y ~ PC + FSL + RUN + SPECIES	0.93	0.6	0.95	2.0
Y ~ FSL + RUN + SPECIES	0.93	0	0.95	0
Y ~ FTL + RUN + SPECIES	0.38	571	0.42	663
Y ~ FSL + SPECIES	0.53	497	0.69	493
Y ~ FSL + RUN	0.58	430	0.54	566

Table 2. Variation in species abundance in the mock-community experiment explained by alternative models. In all models, the response variable Y is the log-transformed amount of DNA used for the focal species as input, and the columns show the proportion of explained variance R^2 and ΔAIC (relative to the model with lowest AIC).

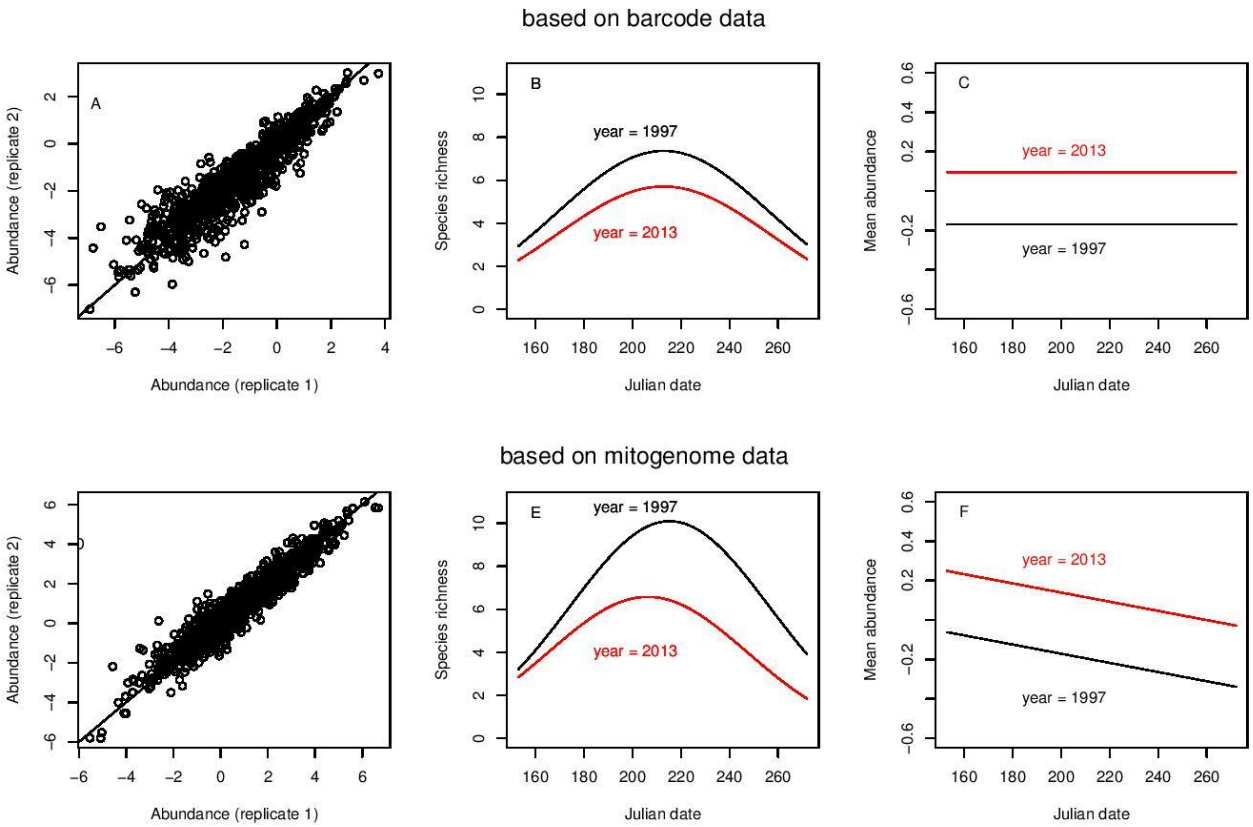
	Barcode	Mitogenome
Number of comparisons	99470	75460
Present in neither run (agreement)	98072	73726
Present in both runs (agreement)	1118	1445
Present in one run (disagreement)	280 (20% of all presences)	289 (17% of presences)

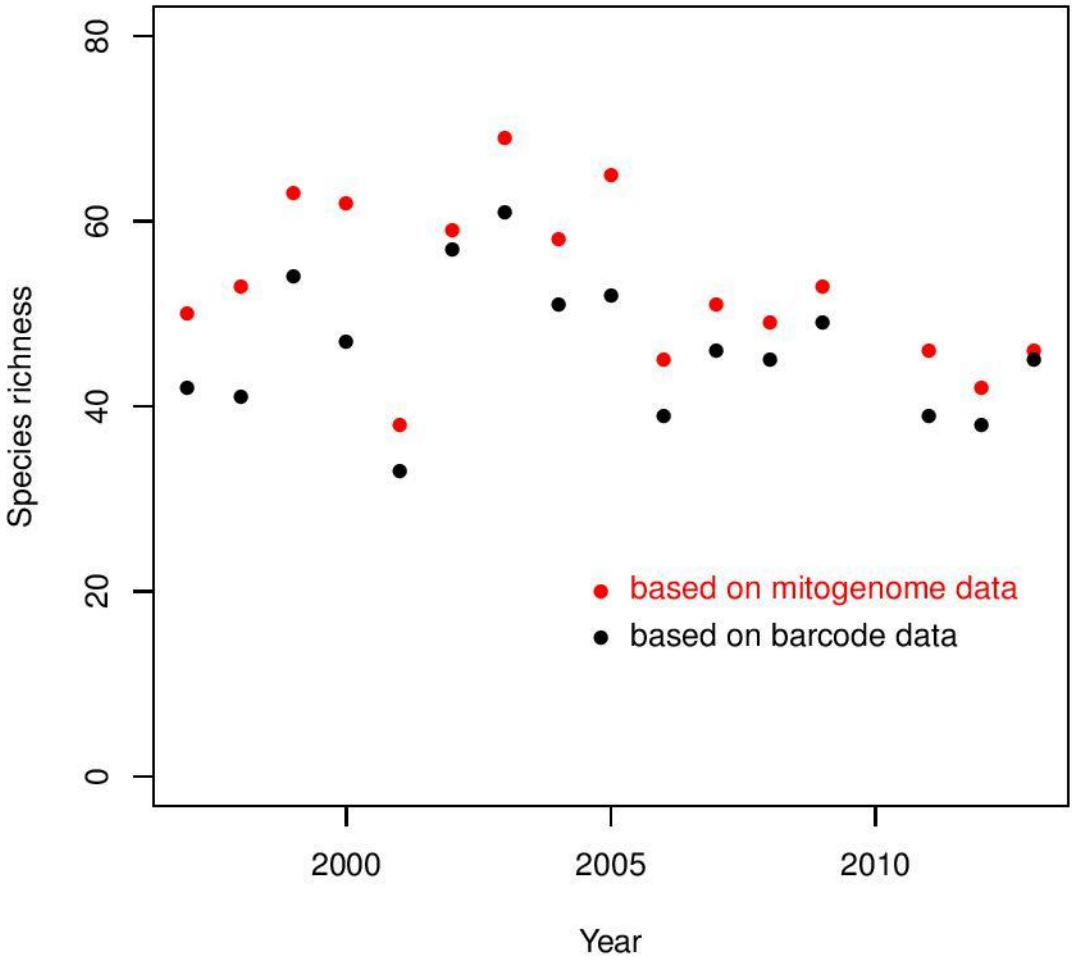
Table 3. Consistency of re-sequenced environmental data. The table shows the consistency in species presences and absences between two sequencing runs, with the total number of comparisons equaling the total number of sample-by-species combinations. To evaluate the robustness of detecting species presence, we counted how often a species was inferred to be present or absent in both cases (agreement) or present in one case only (disagreement). The proportion of presences is counted as the fraction of all samples where the species was inferred to be present in at least one run.

Figure legends

Figure 1. Results from the environmental Zackenberg arthropod samples from 1997 and 2013, collected with yellow pitfall traps and processed using the SPIKEPIPE pipeline. The upper panels (A-C) are based on mapping against DNA barcodes, and the lower panels (D-F) on mapping against mitogenomes. Panels A, D examine the technical validity of the data by comparing run-corrected abundance estimates for species sequenced in technical replicates across sequencing runs. Panels B, E show the species richness per trap-week in 1997 (black line) and 2013 (red line), as predicted by a Poisson regression model fitted to weekly data on all years 1997–2013, as described in the main text. (For clarity, only the starting and end years are shown in the figure.) Panels C, F show the mean abundances per week (in the unit of log-transformed DNA amount), normalized to zero mean and unit variance within each species) of those species that were present in 1997 and in 2013, as predicted by linear regression models described in the main text.

Figure 2. Total number of species detected per year in the Zackenberg samples, based on mapping to either DNA barcodes (black dots) or to mitogenomes (red dots).





Box 1