Highly comparable metabarcoding results from MGI-Tech and Illumina sequencing platforms

Sten Anslan[1,2], Vladimir Mikryukov[1,2], Kęstutis Armolaitis[3], Jelena Ankuda[3], Dagnija Lazdina[4], Kristaps Makovskis[4], Lars Vesterdal[5], Inger Kappel Schmidt[5] and Leho Tedersoo[1,2]

[1] Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Tartumaa, Estonia
[2] Mycology and Microbiology Center, University of Tartu, Tartu, Tartumaa, Estonia
[3] Department of Ecology, Institute of Forestry of Lithuanian Research Centre for Agriculture and Forestry (LAMMC), Kaunas, Lithuania
[4] Latvian State Forest Research Institute SILAVA, Riga, Latvia
[5] Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark

ABSTRACT

With the developments in DNA nanoball sequencing technologies and the emergence of new platforms, there is an increasing interest in their performance in comparison with the widely used sequencing-by-synthesis methods. Here, we test the consistency of metabarcoding results from DNBSEQ-G400RS (DNA nanoball sequencing platform by MGI-Tech) and NovaSeq 6000 (sequencing-by-synthesis platform by Illumina) platforms using technical replicates of DNA libraries that consist of COI gene amplicons from 120 soil DNA samples. By subjecting raw sequencing data from both platforms to a uniform bioinformatics processing, we found that the proportion of high-quality reads passing through the filtering steps was similar in both datasets. Per-sample operational taxonomic unit (OTU) and amplicon sequence variant (ASV) richness patterns were highly correlated, but sequencing data from DNBSEQ-G400RS harbored a higher number of OTUs. This may be related to the lower dominance of most common OTUs in DNBSEQ data set (thus revealing higher richness by detecting rare taxa) and/or to a lower effective read quality leading to generation of spurious OTUs. However, there was no statistical difference in the ASV and post-clustered ASV richness between platforms, suggesting that additional denoising step in the ASV workflow had effectively removed the ‘noisy’ reads. Both OTU-based and ASV-based composition were strongly correlated between the sequencing platforms, with essentially interchangeable results. Therefore, we conclude that DNBSEQ-G400RS and NovaSeq 6000 are both equally efficient high-throughput sequencing platforms to be utilized in studies aiming to apply the metabarcoding approach, but the main benefit of the former is related to lower sequencing cost.

Subjects Bioinformatics, Entomology, Genomics, Molecular Biology, Zoology
Keywords Metabarcoding, COI, Illumina, NovaSeq, DNBSEQ, MGI-Tech

Submitted 2 July 2021
Accepted 14 September 2021
Published 30 September 2021

Corresponding author
Sten Anslan, sten.anslan@ut.ee

Academic editor
Vladimir Uversky

Additional Information and Declarations can be found on page 16

DOI **10.7717/peerj.12254**

How to cite this article Anslan S, Mikryukov V, Armolaitis K, Ankuda J, Lazdina D, Makovskis K, Vesterdal L, Schmidt IK, Tedersoo L. 2021. Highly comparable metabarcoding results from MGI-Tech and Illumina sequencing platforms. **PeerJ 9:e12254**
DOI 10.7717/peerj.12254

INTRODUCTION

Metabarcoding, the identification of organisms *via* DNA marker genes from environmental samples or a mixture of heterospecific specimens (*Taberlet et al., 2018*), is a powerful tool in biodiversity analysis (*Kelly et al., 2018*; *Pont et al., 2021*; *Valentin et al., 2019*; *Watts et al., 2019*). This approach has been efficiently used to characterize the community composition of microbial and animal taxa from various types of environmental samples such as soil (*Bahram et al., 2018*; *Nilsson et al., 2019*), water (*Djurhuus et al., 2018*; *Liu et al., 2020*), sediments (*Kang et al., 2021*; *Wurzbacher et al., 2017*), dust (*de Groot et al., 2021*; *Rocchi et al., 2017*) and feces (*Ando et al., 2020*; *Anslan et al., 2021*). In animals, metabarcoding has also been widely used to identify host-associated microbiomes, determine the structure of entire holobionts and dietary differences in various species (*Alberdi et al., 2019*; *Kueneman et al., 2019*). The information acquired through DNA marker gene sequencing has greatly boosted our knowledge about the ecology and distribution patterns of various aquatic and terrestrial animal groups such as nematodes, arthropods and annelids (*Arribas et al., 2016*; *Beng & Corlett, 2020*; *Compson et al., 2020*; *Deiner et al., 2017*; *Zawierucha et al., 2021*).

Since the mid-2000s, the metabarcoding technique has greatly benefited from technological advances in library preparation, primer and sample-specific index design, novel sequencing platforms as well as from optimized bioinformatics workflows and accumulating reference data (*Taberlet et al., 2018*; *Nilsson et al., 2019*). Short-read, second-generation high-throughput sequencing (HTS) technologies are currently the most widely used means for metabarcoding due to a relatively low cost per sample, high sequencing depth and accuracy. Sequencing instruments produced by Illumina, Inc. (*e.g.*, MiSeq and NovaSeq) using sequencing-by-synthesis technology are dominating the market as they offer viable solutions for both ultra-high sequencing depth and paired-end sequencing of short- and mid-sized amplicons (up to 500–600 bases; *Kumar, Cowley & Davis, 2019*). By utilizing recent advances in DNA nanoball sequencing technology (*Drmanac et al., 2010*; *Li et al., 2019*), MGI-Tech, Inc. has produced several DNBSEQ (MGISEQ) platforms with similar throughput and quality profiles compared with Illumina sequencing (*Jeon et al., 2021*; *Kumar, Cowley & Davis, 2019*). The results from Illumina and MGI-Tech sequencing platforms are highly comparable and may be used interchangeably for RNA sequencing and whole genome sequencing (*Jeon et al., 2019*; *Kim et al., 2021*; *Korostin et al., 2020*). However, the error rate of DNBSEQ technology (MGI-2000 instrument) was marginally higher than for Illumina (HiSeq instrument) when using 2 × 150 paired-end sequencing mode on both platforms (quality scores >30: 95.03% and 97.18% for MGISEQ-2000 and HiSeq 2500, respectively; *Korostin et al., 2020*). The results of these early genome sequencing-oriented studies suggest that MGI-Tech platforms may be used efficiently also in metabarcoding studies. In early 2021, sequencing costs for MGI-Tech DNBSEQ-T7 were about 50% lower compared with Illumina NovaSeq platform (cost per read) for the greatest throughput analyses (*Tedersoo et al., 2021*). So far, only a single metabarcoding study has been conducted to compare these sequencing platforms (DNBSEQ-G400 and Illumina MiSeq) for recovering rRNA gene 16S and

ITS amplicons of bacterial and fungal mock communities (*Sun et al., 2021*). For the ITS2 amplicon, *Sun et al. (2021)* reported small but significant differences between DNBSEQ-G400 and MiSeq platforms, but this difference can be attributed to use of different primer pairs for DNA library preparation.

Here, we aim to compare the relative performance of DNBSEQ-G400RS (2 × 200 bp) and Illumina NovaSeq 6000 (2 × 250 bp) for DNA metabarcoding. DNA libraries were prepared from the same pools of mitochondrial cytochrome oxidase 1 (COI) amplicons generated from soil DNA extracts.

METHODS

Sampling

We selected 120 sites (Table S1) in an area of 500 km^2 around Tartu, Estonia, which included various terrestrial ecosystems (managed croplands, abandoned croplands, plantations of fruit trees and forestry trees on former agricultural land and old forests (>80 years)). In each site, we established a sampling plot (30 × 30 m) in a homogeneous patch to minimize any edge effects and vegetation gradient effects. In a 3 × 3 grid, we collected nine soil cores from each plot, with coordinates representing the central location (Table S1). Soil cores for analyses of soil bulk density were collected by hammering a PVC tube (50 mm diam.) to 100 mm depth after removing loose litter. Using a clean knife (sterilized in 1% NaOCl solution), roughly 15 g of soil was scraped from sides of the same core holes and pooled into a clean Zip-Lock plastic bag. The composite sample was well mixed and frozen immediately among freezing tablets (initial temperature, −86 °C). In the laboratory, the frozen samples were transferred to −80 °C.

Molecular analyses

The frozen samples were crushed using a hammer in double plastic bags (to speed up thawing), placed into paper bags and dried in a drying cabinet at 35 °C for 24 h. The dried samples were transferred into Zip-Lock bags, followed by initial homogenisation by vigorous rubbing by hands. Roughly 1 g of homogenised soil dust was transferred into an Eppendorf tube and subjected to further homogenisation by using two 3-mm steel balls at 30 Hz. Altogether 0.25 g of soil powder was subjected to DNA extraction using a Thermo Scientific KingFisher Flex robot and MagAttract PowerSoil kit (Qiagen Inc., Hilden, Germany), following the manufacturer's instructions.

For amplification *via* PCR, primers mlCOIintF (5′ GGW ACW GGW TGA ACW GTW TAY CCY CC; *Leray et al., 2013*) and jgHCO2198 (3′ TAI ACY TCI GGR TGI CCR AAR AAY CA; *Geller et al., 2013*) were used to target the ~313 bp mitochondrial cytochrome oxidase 1 (COI) gene. The former primer was barcoded with unique phase shift indexes (Table S2). The 25 µl PCR mixture comprised five µl of 5× HOT FIREPol Blend Master Mix (Solis Biodyne, Tartu, Estonia), 0.5 µl of each forward and reverse primer (20 mM), one µl of DNA extract and 18 µl ddH_2O. Thermal cycling included an initial denaturation at 95 °C for 15 min; 25 cycles of denaturation for 30 s at 95 °C, annealing for 30 s at 57 °C, elongation for 1 min at 72 °C; final elongation at 72 °C for 10 min and storage at 4 °C. All PCR reactions were performed in duplicate and pooled for subsequent analyses.

Table 1 **Cost calculations for Illumina NovaSeq 6000 and MGI-Tech DNBSEQ-G400RS based on the best offering service providers and data retrieved (euros).**

	NovaSeq 6000 (2 × 250 bp)	DNBSEQ-G400RS* (2 × 200 bp)
Library preparation for sequencing	100	170
Offer for sequencing 50 million reads	1,000	170
Actual cost per million raw reads	30.07	7.21
Actual cost per million filtered reads (matrix#1)	53.23	11.92
Actual cost per raw gigabit (Gb)	26.44	8.25
Actual cost per filtered Gb (merged and quality filtered)	104.76	33.66

Note:

* Originally in USD; converted to EUR as of 2021-05-10 (invoice issued).

PCR products (five µl) were verified using 1% agarose gel electrophoresis. Samples yielding no product were re-amplified with 30 cycles, followed by pooling and gel electrophoresis. Both a negative control (ddH_2O with no DNA template) and a positive control sample (an artificial DNA molecule with multiple primer sites) were used to assess obvious contamination during sample preparation for PCR and the efficiency of PCR, respectively. The PCR products were normalized for preparation of two libraries based on visual inspection of band strength on an 1% agarose gel. We used the following criteria: no band (*i.e.*, negative control) = 10 µl; faint band = seven µl; medium band = three µl; strong band = one µl. The pooled amplicons were shipped for Illumina NovaSeq 6000 (2 × 250 bp; hereafter NovaSeq) paired-end sequencing in Novogen Inc., UK and for MGI-Tech DNBSEQ-G400RS (2 × 200 bp; hereafter DNBSEQ) paired-end sequencing in Clinomics Inc., South Korea. The service providers were selected strictly based on the best price offer for delivering 50 million reads (Table 1). Sequencing libraries for respective sequencing platforms (including adapter ligation) were prepared by the service providers from the same amplicon pool. NEBNext® Ultra^{TM} II DNA Library Prep Kit (PCR-free workflow) was used for NovaSeq (at Novogene Inc., Cambridge, UK) and MGIEasy FS DNA Library Prep Set (includes PCR step after ligation) was used for DNBSEQ library preparation (at Clinomics Inc., Ulsan, South Korea). The sequencing costs are provided in Table 1.

Bioinformatics

NovaSeq and DNBSEQ provided 42,990,088 and 55,581,045 raw reads, respectively. The raw reads were demultiplexed using cutadapt v3.4 (*Martin, 2011*) by requiring full-length index coverage (–overlap 12) and allowing one mismatch to index sequence (–e 1) but no indels (–no-indels) (Data 1). Fasta-formatted index file specifying sample-index combinations served as an input for cutadapt. The reads were separated into sample-wise files according to this information. During demultiplexing, we also accounted for reverse complementary sequences in the raw data by running two rounds of demultiplexing using cutadapt. For the second round, the unassigned R1 and R2 reads from the first round used as inputs. Demultiplexed reads from both runs were then merged by sample. Prior to further processing, all sequences were reoriented to 5′–3′ orientation based on the PCR primers. For this procedure, fqgrep (v0.4.4; *Das, 2011*) was used by allowing two mismatches to primer sequences as implemented in PipeCraft v1.0

(*Anslan et al., 2017*). Paired-end sequences were assembled using vsearch v2.17.0 (*Rognes et al., 2016*) with the following settings: –fastq_minovlen 10, –fastq_minmergelen 10, –fastq_maxdiffs 20, –fastq_maxns 0, –fastq_maxmergelen 600, –fastq_allowmergestagger. Both forward and reverse primers were trimmed from the sequences using cutadapt by allowing two mismatches to primer strings and primer match overlap of 24 but no indels. Reads where both primers remained undetected were discarded. Quality-filtering of the remaining sequences was performed using trimmomatic v0.39 (*Bolger, Lohse & Usadel, 2014*) with the following options: SLIDINGWINDOW:5:30, LEADING 11, TRAILING 11. Putative chimeric sequences were removed using the *uchime_denovo* algorithm as implemented in vsearch (–id 0.97 for pre-clustering, default options for chimera detection). All filtered sequences from both sequencing platforms per sample were merged and clustered into operational taxonomic units (OTUs) with 97% sequence similarity threshold using vsearch (–cluster_size –iddef 2). During the latter process, a uniform sample-by-OTU table (containing data from both platforms) was generated (matrix#1, see below). The resulting OTUs were classified using BLAST+ v2.10.1 (*Camacho et al., 2009*) against the CO1Classifier database v4 (*Porter & Hajibabaei, 2018*) (Data 2).

In addition to the OTU workflow, amplicon sequence variants (ASVs) were calculated using DADA2 (v1.18; *Callahan et al., 2016*). The ASV matrix was generated including a subset of 60 samples from the total of 120 samples that were used for generating OTU matrices (Table S1). The quality filtering options included removal of all sequences with ambiguous bases (maxN = 0), trimming low quality ends (truncQ = 2) and keeping sequences with maximum expected error rates of one (maxEE = 1). Chimeras were removed with the 'consensus' method. All other processes, including denoising, followed the default DADA2 workflow (Data 1). Inputs for the ASV pipeline included fastq files for each sample that comprised primer-trimmed and reoriented reads (as specified above). The ASVs were further subjected to post-clustering at 97% sequence similarity threshold using the LULU algorithm (*Frøslev et al., 2017*) to merge consistently co-occurring AS*V*s. Thus, two sets of ASVs matrices were generated for the analyses: (1) ASVs matrix, and (2) post-clustered ASVs matrix (Data 3).

OTU data matrices

Four types of sample-by-OTU data matrices were prepared (Data 1) to compare the DNBSEQ and NovaSeq sequencing platforms: (1) matrix#1—all 'raw' OTUs as outputted after the clustering step; (2) matrix#2—'raw' OTUs but global singletons (*i.e*., OTUs that had only one sequence across matrix#1) removed; (3) matrix#3—only metazoan global non-singleton OTUs; (4) matrix#4—rarefied metazoan OTUs. To account for variations in sequencing depth, metazoan OTUs data (matrix#3) was rarefied using phyloseq v1.34.0 (*McMurdie & Holmes, 2013*) to a depth of 10,489 reads per sample (matrix#4). To reduce the remaining putative artefacts in the matrix#3, OTUs with a representative sequence length different from the expected amplicon length (313 bp ± 4 bp) were discarded. This eliminated 59% (82,036) of OTUs accounting for 31.8% (15,639,811) of reads from matrix#2. Besides Metazoa, the used COI primers amplified a wide variety of other non-target eukaryotes (mostly fungi) as well as prokaryotes. An OTU was assigned to

Metazoa (in matrix#3) when the best blastn match of the query sequence had at least 90% query coverage and 75% identity against the reference sequence. For metazoan taxonomic group statistics, an OTU received a phylum level classification when the best blastn match of the query sequence (an OTU) had ≥80% identity against the reference sequence with phylum-level annotation. Some OTUs were best matched to Hydrozoa and Porifera at <89% sequence similarity, but since these aquatic organisms are unexpected in terrestrial environments, we assigned these OTUs to unclassified Metazoa.

Statistics

Differences in the OTU/ASV richness between DNBSEQ and NovaSeq data sets were tested using paired *t*-tests in STATISTICA (v7; StatSoft Inc., Tulsa, OK, USA). For ASV matrices and OTU matrices#1–3, we first calculated the predicted richness values based on residuals of OTU richness, as derived from linear regression analyses using natural logarithm transformed sequencing depth as an independent variable, separately for DNBSEQ and NovaSeq data subsets. For the OTU matrix#4, residuals were not calculated because of using rarefaction. Spearman correlation was used to examine the sequencing depth and OTU richness correlations between sequencing platforms. Mantel tests (with 9999 permutations, method = 'spear'), as implemented in the 'vegan' package v2.5.7 (*Oksanen et al., 2015*) in R v4.0.4 (*R-Core-Team, 2021*), were used to test correlations between corresponding sample similarities from different sequencing platforms. Additionally, Procrustes tests (with 9999 permutations, metaMDS ordination), as implemented in the 'vegan' package, were used to compare correlation in community structure as revealed from DNBSEQ and NovaSeq instruments. Bray–Curtis similarity of Hellinger-transformed data were used for both Mantel and Procrustes tests. To assess OTU/ASV overlap between sequencing platforms, Venn diagrams were drawn using Venny 2.1 (*Oliveros, 2018*). The proportion of potential index-switching errors was estimated using the UNCROSS2 score (*Edgar, 2018*) with default parameter values ($f = 0.01$, $t_{min} = 0.1$) for each sample and sequencing platform combination. Differences among sequencing platforms were tested using a Bayesian generalized linear mixed model with binomial errors and logit link, where 'sample' was used as a random effect. The model was fitted with Stan v2.21 (*Stan-Development-Team, 2021*) and brms package v2.15.0 (*Bürkner, 2017*) using seven Markov chains of Hamiltonian Monte Carlo, with 15,000 sampling iterations and 2,000 warm-up iterations for each chain.

RESULTS

Demultiplexed HTS datasets of 120 samples from DNBSEQ and NovaSeq contained 50,129,600 and 39,813,707 sequences, respectively. The overall quality score distributions exhibited similar profiles between DNBSEQ and NovaSeq datasets (Fig. 1; Fig. S1). However, the latter exhibited marginally higher level of expected number of errors in the sequences (Fig. 1A). Therefore, after filtering (all filtering steps before clustering in the OTU workflow), proportionally more sequences were discarded in the NovaSeq data (48.1%) compared with DNBSEQ data (43.1%; Table S3). Similarly, after the ASVs

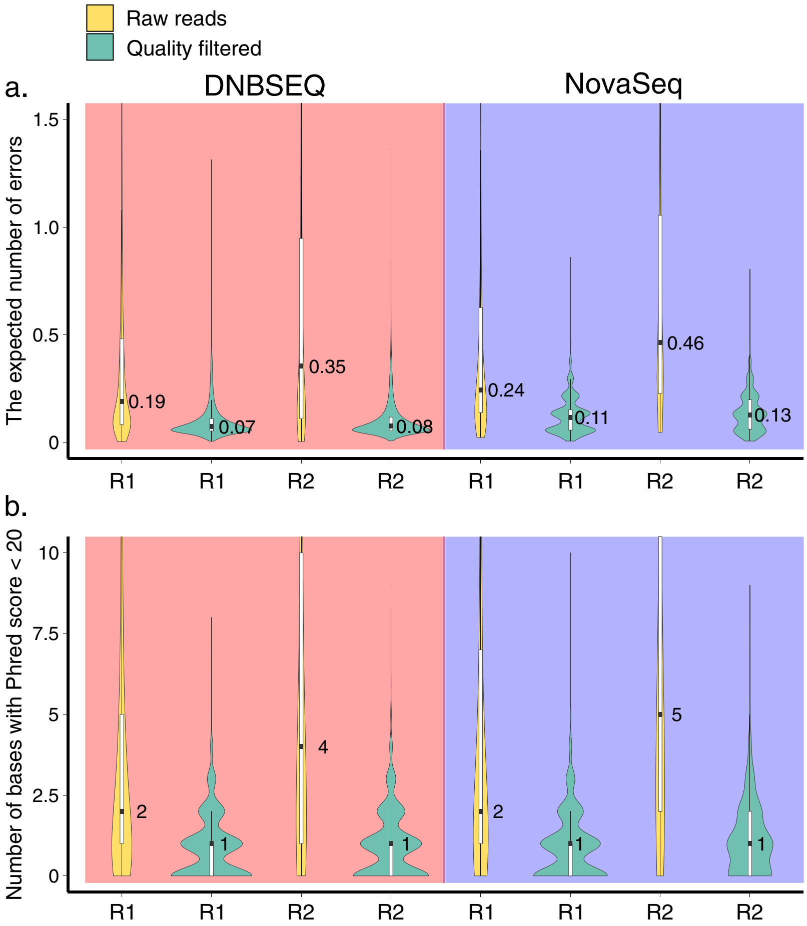

Figure 1 **Quality profile.** Distributions of the expected number of errors (A) and the number of bases with Phred score <20 (B) per read in raw (yellow) and trimmomatic quality-trimmed (green) reads obtained with DNBSEQ and NovaSeq platforms for a single sample (G5300; 392,379 and 270,651 raw reads in DNBSEQ and NovaSeq data, respectively). Median, quartiles, and the 1.5 interquartile range are shown on the boxplots inside the violin plots. Due to the highly skewed distributions, *y*-axes were truncated to 1.5 (A) and 10 (B). The expected number of errors per sequence was estimated following *Edgar & Flyvbjerg (2015)*.

Full-size DOI: 10.7717/peerj.12254/fig-1

workflow (for the subset of 60 samples out of 120), the average proportional sequence loss per sample was higher in NovaSeq data (29.5% *vs*. 33.4%; Table S3).

Clustering at 97% sequence similarity threshold (both datasets merged) revealed 182,066 OTUs including 43,136 singletons. Of the 138,930 non-singleton OTUs, 17,547 (12.6%) were unique to DNBSEQ and 20,175 (14.5%) unique to NovaSeq (Fig. 2A). These unique OTUs usually comprised a low number of reads, with a median sequence count of 3 (±13.7 SD) and 5 (±61.1 SD) for DNBSEQ and NovaSeq data, respectively. The proportion of shared OTUs between datasets was 72.8% for all non-singleton OTUs (matrix#2), 96.6% for the full metazoan dataset (matrix#3) and 85.2% for the rarefied metazoan dataset (matrix#4; Fig. 2). Total number of ASVs (in a subset of 60 samples, across both data sets) was 121,402, including 2,660 global singletons. Post-clustering those ASVs with 97% sequence similarity threshold with LULU algorithm merged 17,756 ASVs (14.6%),

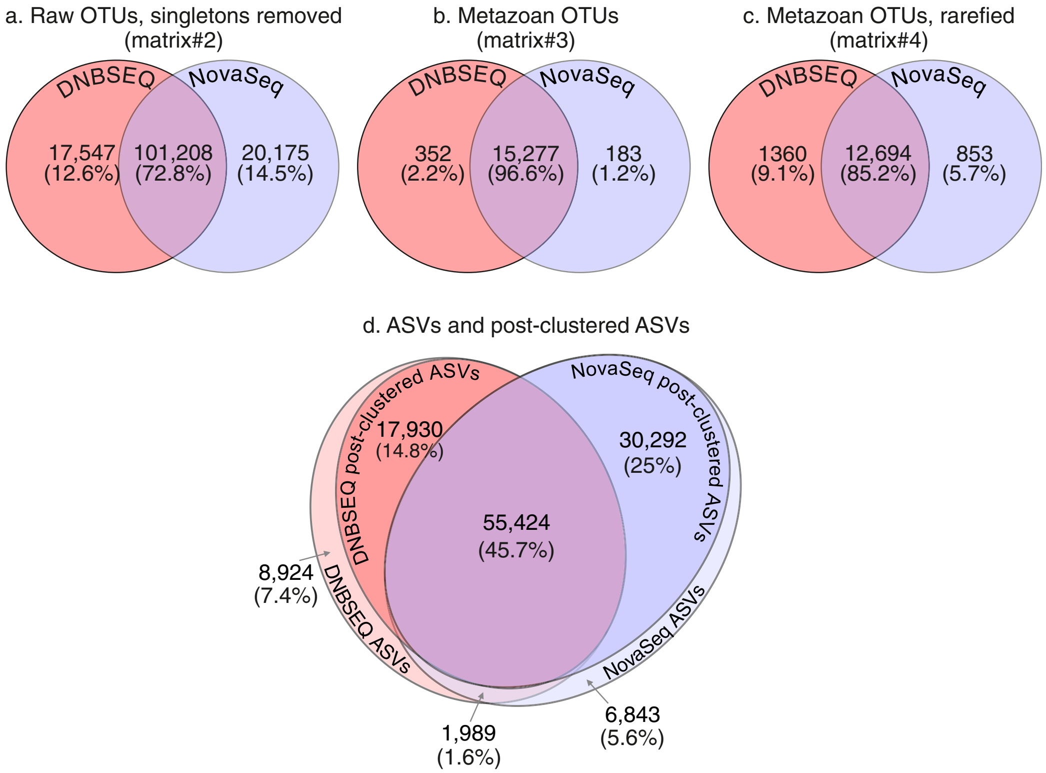

Figure 2 Venn diagrams. Venn diagrams demonstrating the number and proportions of shared and unique OTUs per sequencing platform: (A) raw OTUs, global singletons excluded (matrix#2); (B) metazoan OTUs, global singletons excluded (matrix#3); (C) rarefied metazoan OTUs (matrix#4); (D) ASVs and post-clustered ASVs matrices (note that ASVs data is for a subset of 60 samples). The total proportion of shared ASVs is 47.3%, but 53.5% when considering only post-clustered ASVs.

Full-size DOI: 10.7717/peerj.12254/fig-2

retaining 103,646 ASVs (including global 343 singletons). The proportion of shared ASVs between post-clustered ASVs data set was 53.5% (Fig. 2D). The unique post-clustered ASVs per platform had median sequence count of 7 (±56.9 SD) and 5 (±42.0 SD) for DNBSEQ and NovaSeq data, respectively.

Taxonomic composition

Metazoa contributed to 13.2% and 12.7% OTU richness, and 18.6% and 24.4% total read abundance, in the DNBSEQ and NovaSeq datasets, respectively. Within Metazoa, the largest phyla in terms of OTU richness were Arthropoda and Nematoda, whereas the largest classes were Insecta (Arthropoda) and Chromadorea (Nematoda) (Fig. 3). While the distribution of relative OTU numbers was highly similar across sequencing platforms, there were certain differences in relative abundance of reads. In the DNBSEQ data, relatively more reads of unclassified Metazoa were recovered at the expense of Annelida (Fig. 3). Taxonomic annotation of ASVs were not performed in this study.

Sequencing depth and diversity

While the total sequencing depth among platforms depended on the sequencing amount ordered and cannot be thus compared, the relative proportion of reads per sample was

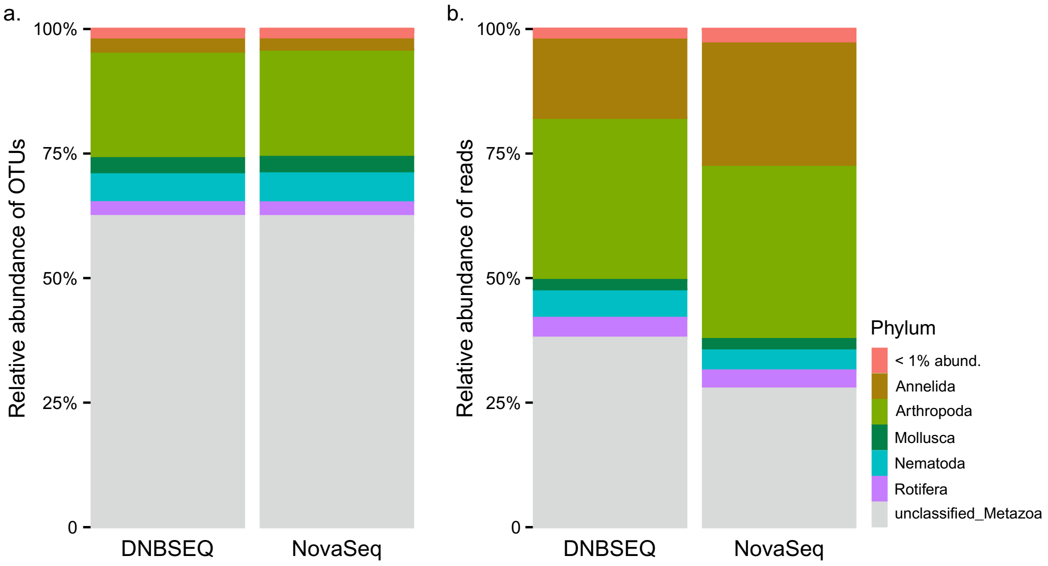

Figure 3 **Taxonomy.** Phylum-level bar plots indicating proportional distribution of metazoan OTUs (A) and sequences (B) from DNABSEQ and NovaSeq platforms (data matrix#3).

Full-size DOI: 10.7717/peerj.12254/fig-3

highly similar across the two platforms (min: 0.18% and 0.18%; max: 2.57% and 2.66%; median: 0.73% and 0.73%; average: 0.83% and 0.83% for DNBSEQ and NovaSeq, respectively; Fig. S2). All four OTU matrices exhibited strong correlations in per-sample OTU richness between sequencing platforms (Fig. 4A), with Spearman correlation coefficients 0.974, 0.974, 0.994 and 0.970 for raw OTUs (matrix#1), non-singleton OTUs (matrix#2), metazoan OTUs (matrix#3) and rarefied metazoan OTUs (matrix#4), respectively (P < 0.001 for all tests). Similarly, all OTU matrices exhibited strong correlations in per-sample sequence abundance between sequencing platforms (Fig. 4B), with Spearman correlation coefficients 0.975, 0.975 and 0.912 for matrix#1, matrix#2 and matrix#3, respectively (P < 0.001 for all cases). In addition, community composition retrieved by DNBSEQ and NovaSeq platforms were strongly correlated based on Procrustes ($R \geq 0.97$ and $P < 0.001$ for all tests) and Mantel (mantel $R \geq 0.991$ and $P < 0.001$ for all tests) statistics (Fig. 5). These patterns were the nearly identical when comparing ASV matrices of DNBSEQ and NovaSeq (Fig. 5; Fig. S3).

There was a significant difference in OTU richness between DNBSEQ and NovaSeq data in all OTU matrices (matrix#1: t = 39.191, df = 119, P < 0.001; matrix#2: t = 40.140, df = 119, P < 0.001; matrix#3: t = 15.755, df = 119, P < 0.001; and matrix#4: t = 22.723, df = 119, P < 0.001) (Figs. 6A–6C). For example, an average per-sample OTU richness was 9.7% higher in DNBSEQ data in the rarefied metazoan dataset (matrix#4). The rank abundance curves of OTUs (matrix#4) derived from DNBSEQ and NovaSeq displayed a slightly different pattern (Fig. 7). There was a slight tendency towards greater dominance in the NovaSeq dataset, with the top three abundant OTUs being more abundant by a factor of 2.9, 3.2 and 1.9. We further explored the differences in OTU richness between data sets by removing all potential 'noise' of spurious OTUs by further filtering matrix#4

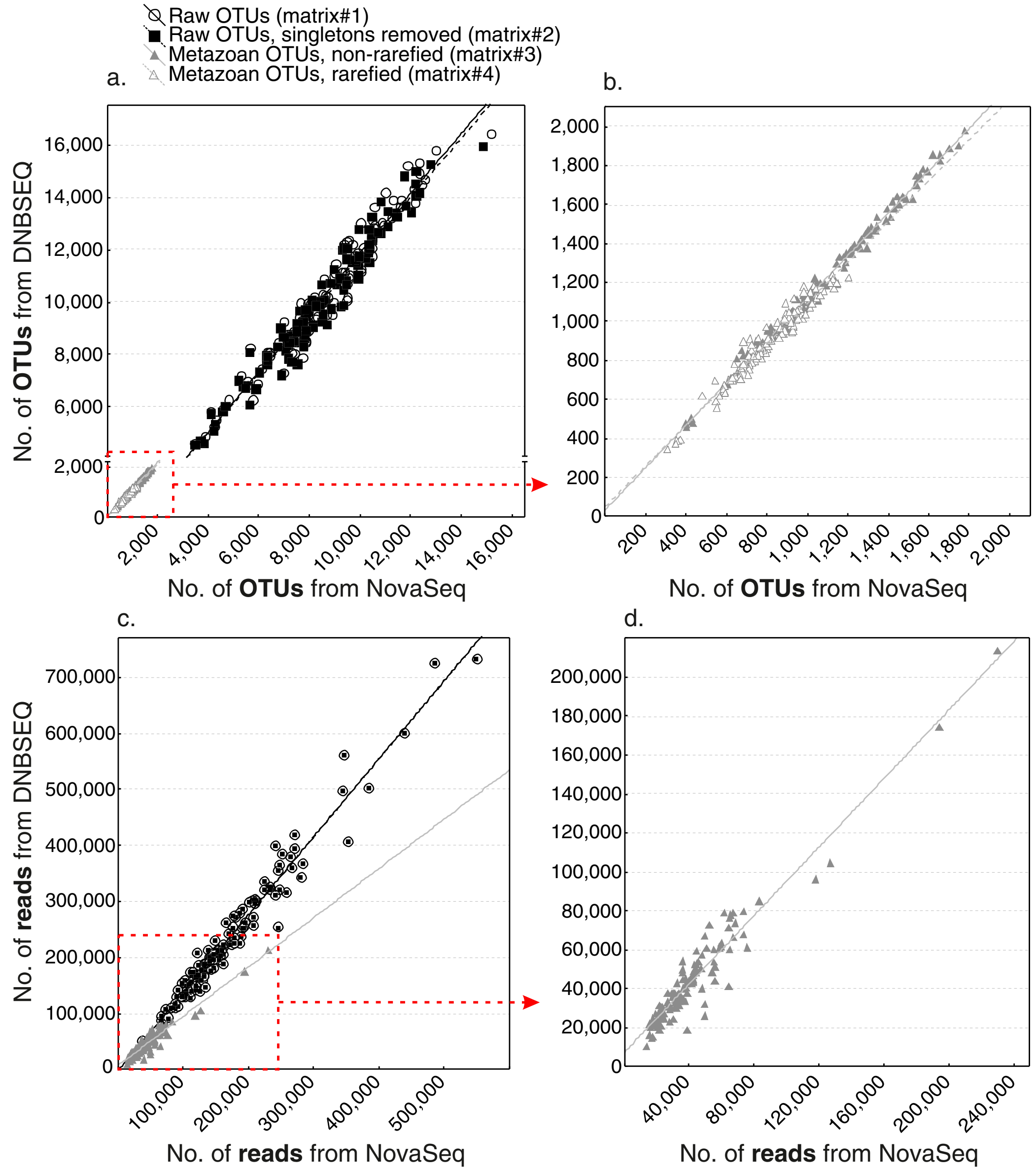

Figure 4 Richness correlations. Correlation of OTU richness (A and B) and read abundance (C and D) per sample as provided by DNBSEQ and NovaSeq platforms (Spearman $R \geq 0.91$, $N = 120$, $P < 0.001$ for all matrices using logarithm-transformed data). Plot b. represents a zoom-in for OTU matrix#3 and #4. Plot d. is a zoom-in for matrix#3. Note that the differences of number of reads from matrix#1 and matrix#2 are relatively low, thus the points completely overlap in the low-resolution plot c.

Full-size DOI: 10.7717/peerj.12254/fig-4

(rarefied metazoan OTU table) to include only OTUs with relative sequence abundance of ≥0.01% (per data set) and ≥98% sequence similarity to the reference sequences (matrix#5 in Data 1). In this stringently filtered data set, differences in OTU richness disappeared (paired *t*-test: $t = 0.131$, df = 119, $P = 0.896$; Fig. 6D). Similarly, there were no significant differences in the ASV richness between DNBSEQ and NovaSeq data sets in the ASV matrices ($P > 0.9$; Figs. 6E and 6F).

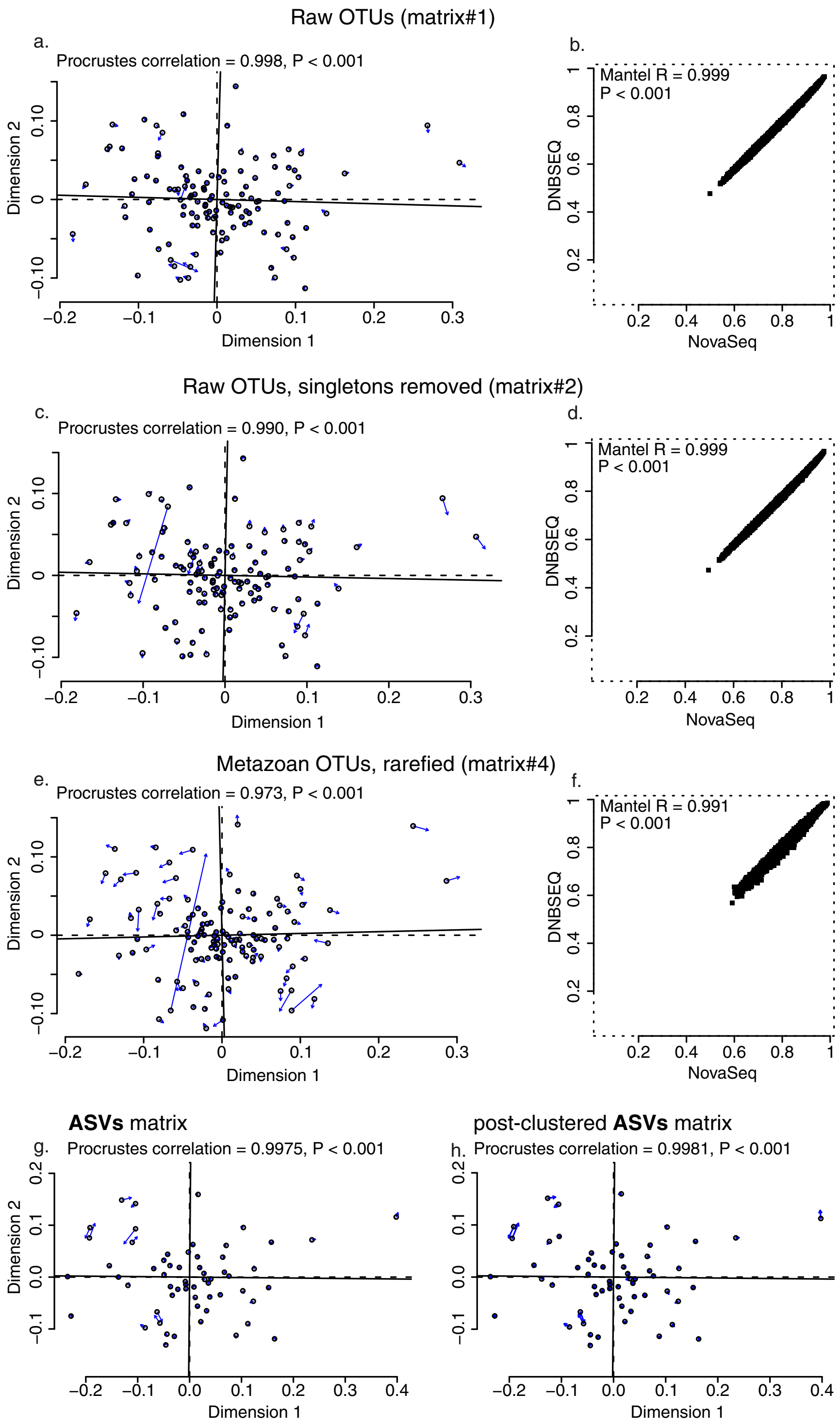

Figure 5 **Community correlations.** Procrustes (A, C and E) and Mantel test plots (B, D and F) for OTU data matrices: (A and B) raw OTUs (matrix#1), (C and D) raw OTUs without singletons (matrix#2), (E and F) metazoan OTUs with rarefied data (matrix#4). Plots for metazoan global non-singleton OTU matrix (matrix#3) not displayed here, but the analyses revealed the same results (significantly high correlation between data sets, Procrustes correlation = 0.995, $P < 0.001$; Mantel $R = 0.996$, $P < 0.001$). Open circles in the Procrustes plots denote ordination configuration of NovaSeq data and arrows point to the configuration in DNBSEQ data ordination. Full-size DOI: 10.7717/peerj.12254/fig-5

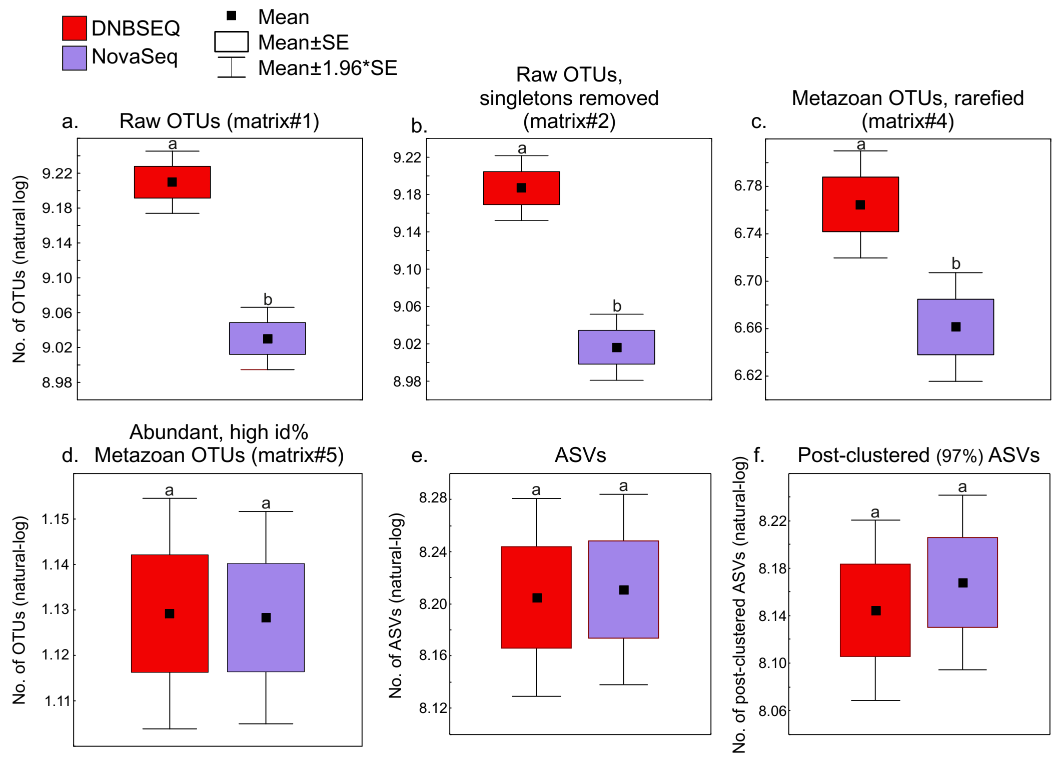

Figure 6 Richness. Box-plots showing platform-wise differences in OTU/ASV richness in (A) raw OTUs (matrix#1), (B) raw OTUs without singletons (matrix#2) and (C) metazoan OTUs (rarefied; matrix#4) richness data. Box-plot for metazoan global non-singleton OTU matrix (matrix#3) not displayed here, but the analyses revealed the same results (significantly higher number of OTUs in the DNBSEQ data set, $P < 0.001$). (D) number of ASVs and (E) post-clustered ASVs. The ASVs dataset includes a subset of 60 samples out of total 120 samples (note that OTU results were the same for corresponding 60 samples as reported here for the 120 samples). Different letters above the whiskers denote statistical differences of the means.

Full-size DOI: 10.7717/peerj.12254/fig-6

Index-switching errors

The UNCROSS2 score revealed that index-switching errors were slightly higher in the DNBSEQ OTU matrices #1 and #2 (Fig. 8; Table S5). For example, the overall proportion of reads that represent putative index switches were 0.049% and 0.038% for DNBSEQ and NovaSeq data in matrix#1, respectively (Fig. 8). However, in the rarefied metazoan dataset (matrix#4), the DNBSEQ matrix displayed a lower proportion of index-switching errors compared with the NovaSeq data (0.021% *vs.* 0.043%; Fig. 8; Table S5). This indicates that rarefaction either lowers the detection of index-switching errors or the majority of index-switches (which occur in low abundances) were removed during the process. However, compared with the OTU matrices, the ASV matrices displayed a relatively lower proportion of reads with putative index-switching errors, and the data from both platforms exhibited similar level of index switches (Fig. 8). Procrustes correlations between index-switch corrected and uncorrected OTU tables were high (0.977–0.999; $P < 0.001$; Fig. S3), indicating that quantitative community-level analyses are weakly impacted by these low proportions of index-switching errors.

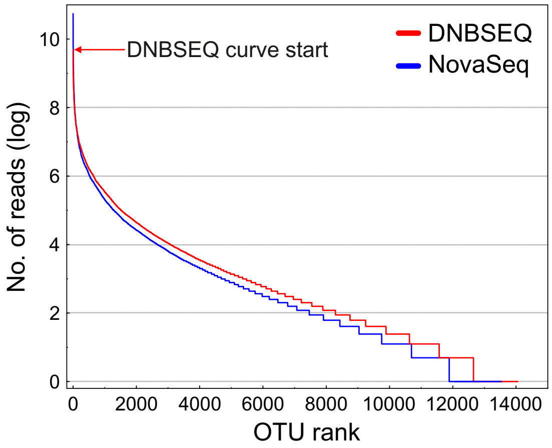

Figure 7 Rank abundance curves (RAC). Rank abundance curves for rarefied metazoan OTUs (matrix#4) based on natural logarithm transformed abundance. The top position of the DNBSEQ curve is marked with a red arrow. The most abundant DNBSEQ OTU ranks seventh overall. The six most abundant NovaSeq OTUs are on average 2.3-fold more abundant than the corresponding DNBSEQ OTUs. Full-size DOI: 10.7717/peerj.12254/fig-7

DISCUSSION

Recovering OTU/ASV richness and composition

By using the same amplicon pools of ~313-bp COI marker gene fragment for platform-specific library preparation and sequence data generation on DNBSEQ-G400RS and NovaSeq 6000 platforms, we demonstrate strongly correlating community and richness profiles. The overall similarities between two short-read sequencing platforms corroborate earlier studies on metabarcoding of bacteria (*Sun et al., 2021*) and genomics of various organisms (*Jeon et al., 2021*; *Kim et al., 2021*).

The OTU and ASV community patterns showed strong correlations between the sequencing platforms (Fig. 5; Fig. S3), but the DNBSEQ dataset revealed on average 9.7% higher OTU richness per sample (rarefied metazoan OTUs, matrix#4). This may be related to a lower effective read quality leading to generation of spurious OTUs (Edgar 2017). To test this, we clustered the unique NovaSeq and DNBSEQ metazoan OTUs in matrix#4 at 96% sequence similarity. Altogether 10.2% and 12.9% of the unique NovaSeq and DNBSEQ OTUs (in matrix#4) clustered to other OTUs using this relaxed threshold. This difference suggests that the greater number of closely related OTUs may result from a slightly higher proportion of remaining erroneous reads in the DNBSEQ data. Furthermore, the DNBSEQ data exhibited lower relative abundance (in terms of number of reads) of the most common OTUs compared with the NovaSeq dataset (Fig. 7), which may also result from a higher proportion of sequencing errors. However, if this greater dominance is artefactual, occupation of a large proportion of sequences by the dominants may render rare species undetected and result in a lower overall richness (*Elbrecht, Peinert & Leese, 2017*). Therefore, less biased sequencing depth towards high abundant taxa may result in overall greater detected richness (*Elbrecht, Peinert & Leese, 2017*). In this study, we did not include a relevant mock community and therefore, we cannot compare

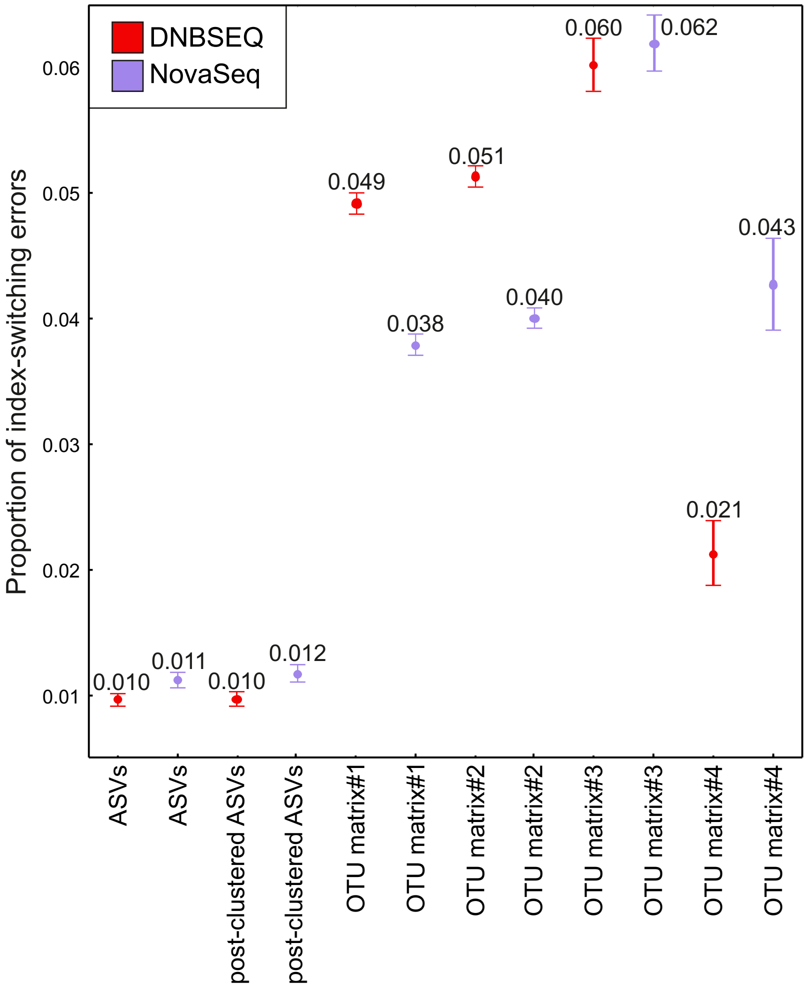

Figure 8 **Index switches.** The proportion of index-switching reads for tested OTU and ASV matrices from DNBSEQ (red) and NovaSeq (blue) sequencing platforms. Error bars denote 95% confidence intervals (see Table S5). *N* = 120 for OTU matrices and *N* = 60 for ASV matrices. The index-switching proportions were similar as displayed here when testing the OTU tables comprising 60 corresponding samples. Full-size DOI: 10.7717/peerj.12254/fig-8

whether the results of one or the other platform are closer to the reality in terms of Metazoan diversity. However, the indications about the higher proportion of remaining erroneous reads in the DNBSEQ data (after quality filtering) was also supported by the disappearance of significant differences in the OTU richness when comparing the stringently filtered matrix#5 (Fig. 6D). Furthermore, the ASV matrices demonstrated highly similar richness profiles between different platforms (Figs. 6E and 6F). The ASV workflow included the DADA2 denoising algorithm (*Callahan et al., 2016*), which seems to efficiently remove the remaining 'noise' and resulting in highly concurrent ASV and post-clustered ASV richness profiles between the sequencing platforms (Figs. 6E and 6F). A majority of the 'noise' is 'hidden' in the molecular units (OTUs or ASVs) with low read count, especially in the sample-wise singletons. Compared with the full ASV matrix (both DNBSEQ and NovaSeq data), the sample-wise singletons were 65 times more

abundant in the OTU matrix#2 (a comparison across 60 samples; Data 1), which contributed to the OTU richness differences between sequencing platforms.
The bioinformatics workflow with the additional denoising step lowers the fraction of low-abundance spurious molecular units, which inflate the richness (*Reitmeier et al., 2021*). Additionally, rare OTUs (*i.e.*, OTUs with low number of reads) are poorly reproducible between sequencing runs (*Leray & Knowlton, 2017*). Therefore, non-stringent quality-filtering may increase richness heterogeneity for the same samples sequenced in different runs (*Reitmeier et al., 2021*). Despite the differences in OTU richness in our study, the OTU community level analyses from either platform would yield highly corresponding results (as indicated by the high Procrustes and Mantel correlations, >0.97; Fig. 5); however, an additional denoising and filtering low abundant molecular units may aid towards more accurate richness analyses.

Index switches

Potential index-switching errors in the raw OTU matrices #1 and #2 were slightly higher in the DNBSEQ than NovaSeq data (Fig. 8; Table S5). This may be at least partly related to the library preparation processes (by service providers) prior to sequencing. The NovaSeq library preparation included a PCR-free workflow, whereas the DNBSEQ library was subjected to post-ligation PCR which may have significant effect on index switching (*Schnell, Bohmann & Gilbert, 2015*; *Caroe & Bohmann, 2020*). While index switches had a negligible impact on the community analyses, such a slightly higher index switch rate in the DNBSEQ data may partly explain the observed differences in per-sample OTU richness (Fig. 6). Following rarefaction (matrix#4), the proportion of potential index-switching errors decreased considerably. This indicates that many potential index-switching errors were removed by discarding a large proportion of sample-wise rare OTUs (with low read abundance), which are more likely to be technical artefacts. Because of higher sequencing depth in the DNBSEQ data in our study (and higher per-sample singleton OTU proportion in matrix#3), latter data set lost proportionally more reads and probably therefore the proportion of putative index switches declined slightly more in the DNBSEQ data (Fig. 8). Because many low abundant sequences (especially singletons) were removed during the ASV workflow, the index switches in the corresponding matrices displayed markedly lower proportion of putative index-switching errors, with a highly similar proportion of index switches remaining in both datasets (Fig. 8). Although index switches are a known issue in high-throughput sequencing platforms (*Carlsen et al., 2012*; *Caroe & Bohmann, 2020*; *Loit et al., 2019*; *Schnell, Bohmann & Gilbert, 2015*), we found that it had a minor effect on the community structure in our tested datasets (Fig. S5). Nonetheless, being aware of the presence of such errors and applying appropriate data curation prior to statistical analyses are principal requisites of a scientific study (*Caroe & Bohmann, 2020*; *Esling, Lejzerowicz & Pawlowski, 2015*).

Limitations

The main limitation of this study is no replication of sequencing runs with companies providing a similar service. However, the single runs of DNBSEQ-G400RS and NovaSeq

6000 revealed similar results, which is unlikely to occur when one or both of these runs have technical issues or biases in library preparation. By testing the reproducibility of 16S amplicon sequencing results from Illumina MiSeq platform, *Wen et al. (2017)* demonstrated that the OTU community variations were greater between technical replicates that were subjected to different sequencing runs compared with variations that were derived from technical replicates within the same sequencing run. Relatively higher variations between different sequencing runs are likely arising because of the low reproducibility of rare OTUs (*i.e.*, OTUs with low number of reads; *Leray & Knowlton, 2017*). Here, we intentionally excluded a mock community because we did not access various axenically grown animals.

CONCLUSIONS

We demonstrate that the MGI-Tech DNBSEQ-G400RS and Illumina NovaSeq 6000 instruments are both well suited for DNA metabarcoding of COI amplicon libraries of ~313 bases given the similarities in data quality and reconstruction of animal diversity. However, we caution that amplicon length (beyond 350 bases) and length heterogeneity (some amplicons beyond 350 bases such as in fungal Internal Transcribed Spacer, ITS) may become critical for the 2 × 200 paired-end chemistry of the MGI-Tech DNBSEQ-G400RS instrument. We conclude that the main benefit of DNA nanoball sequencing lies in its lower sequencing costs (Table 1).

ACKNOWLEDGEMENTS

We thank the service providers for providing excellent-quality data. We have no conflicting interests with the service providers. We thank Vasco Elbrecht, Mathilde Borg Dahl and an anonymous reviewer for their constructive comments.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Funding for this study was provided by the Novo Nordisk Foundation grant NNF20OC0059948 (Silva Nova), EEA Financial Mechanism Baltic Research Programme in Estonia (SUCC, EMP442) and the Estonian Research Council (Mobilitas Pluss grant MOBTP198). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Novo Nordisk Foundation: NNF20OC0059948.
EEA Financial Mechanism Baltic Research Programme in Estonia: SUCC, EMP442.
Estonian Research Council: MOBTP198.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Sten Anslan conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Vladimir Mikryukov performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Kęstutis Armolaitis performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Jelena Ankuda performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Dagnija Lazdina performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Kristaps Makovskis performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Lars Vesterdal performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Inger Kappel Schmidt performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Leho Tedersoo conceived and designed the experiments, performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw sequencing data is available in the Sequence Read Archive: PRJNA743174.

All data matrices generated from the sequencing data and used for the analyses are available in the Supplemental File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.12254#supplemental-information.

REFERENCES

Alberdi A, Aizpurua O, Bohmann K, Gopalakrishnan S, Lynggaard C, Nielsen M, Gilbert MTP. 2019. Promises and pitfalls of using high-throughput sequencing for diet analysis. *Molecular Ecology Resources* **19(2)**:327–348 DOI 10.1111/1755-0998.12960.

Ando H, Mukai H, Komura T, Dewi T, Ando M, Isagi Y. 2020. Methodological trends and perspectives of animal dietary studies by noninvasive fecal DNA metabarcoding. *Environmental DNA* **2(4)**:391–406 DOI 10.1002/edn3.117.

Anslan S, Bahram M, Hiiesalu I, Tedersoo L. 2017. PipeCraft: flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data. *Molecular Ecology Resources* **17(6)**:e234–e240 DOI 10.1111/1755-0998.12692.

Anslan S, Li H, Künzel S, Vences M. 2021. Microbiomes from feces vs. gut in tadpoles: distinct community compositions between substrates and preservation methods. *SALAMANDRA* **57**:96–104.

Arribas P, Andujar C, Hopkins K, Shepherd M, Vogler AP. 2016. Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in Ecology and Evolution* **7(9)**:1071–1081 DOI 10.1111/2041-210X.12557.

Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, Bengtsson-Palme J, Anslan S, Coelho LP, Harend H. 2018. Structure and function of the global topsoil microbiome. *Nature* **560(7717)**:233–237 DOI 10.1038/s41586-018-0386-6.

Beng KC, Corlett RT. 2020. Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. *Biodiversity and Conservation* **29(7)**:2089–2121 DOI 10.1007/s10531-020-01980-0.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30(15)**:2114–2120 DOI 10.1093/bioinformatics/btu170.

Bürkner P-C. 2017. brms: an R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**:1–28.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods* **13(7)**:581–583 DOI 10.1038/nmeth.3869.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10(1)**:421 DOI 10.1186/1471-2105-10-421.

Carlsen T, Aas AB, Lindner D, Vrålstad T, Schumacher T, Kauserud H. 2012. Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecology* **5(6)**:747–749 DOI 10.1016/j.funeco.2012.06.003.

Caroe C, Bohmann K. 2020. Tagsteady: a metabarcoding library preparation protocol to avoid false assignment of sequences to samples. *Molecular Ecology Resources* **20(6)**:1620–1631 DOI 10.1111/1755-0998.13227.

Compson ZG, McClenaghan B, Singer GAC, Fahner NA, Hajibabaei M. 2020. Metabarcoding from microbes to mammals: comprehensive bioassessment on a global scale. *Frontiers in Ecology and Evolution* **8**:192 DOI 10.3389/fevo.2020.581835.

Das I. 2011. fqgrep. *Available at https://github.com/indraniel/fqgrep*.

de Groot GA, Geisen S, Wubs EJ, Meulenbroek L, Laros I, Snoek LB, Lammertsma DR, Hansen LH, Slim PA. 2021. The aerobiome uncovered: multi-marker metabarcoding reveals potential drivers of turn-over in the full microbial community in the air. *Environment International* **154**:106551 DOI 10.1016/j.envint.2021.106551.

Deiner K, Bik HM, Machler E, Seymour M, Lacoursiere-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L. 2017. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Molecular Ecology* **26(21)**:5872–5895 DOI 10.1111/mec.14350.

Djurhuus A, Pitz K, Sawaya NA, Rojas-Marquez J, Michaud B, Montes E, Muller-Karger F, Breitbart M. 2018. Evaluation of marine zooplankton community structure through environmental DNA metabarcoding. *Limnology and Oceanography-Methods* **16(4)**:209–221 DOI 10.1002/lom3.10237.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327(5961)**:78–81 DOI 10.1126/science.1181498.

Edgar RC, Flyvbjerg H. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads, accuracy of microbial community diversity estimated by closed-and open-reference OTUs. *Bioinformatics* **31(21)**:3476–3482 DOI 10.1093/bioinformatics/btv401.

Edgar RC. 2018. UNCROSS2: identification of cross-talk in 16S rRNA OTU tables. *bioRxiv* 400762 DOI 10.1101/400762.

Elbrecht V, Peinert B, Leese F. 2017. Sorting things out: assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution* **7(17)**:6918–6926 DOI 10.1002/ece3.3192.

Esling P, Lejzerowicz F, Pawlowski J. 2015. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research* **43(5)**:2513–2524 DOI 10.1093/nar/gkv107.

Frøslev TG, Kjøller R, Bruun HH, Ejrnæs R, Brunbjerg AK, Pietroni C, Hansen AJ. 2017. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications* **8(1)**:1188 DOI 10.1038/s41467-017-01312-x.

Geller J, Meyer C, Parker M, Hawk H. 2013. Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources* **13(5)**:851–861 DOI 10.1111/1755-0998.12138.

Jeon SA, Park JL, Kim J-H, Kim JH, Kim YS, Kim JC, Kim S-Y. 2019. Comparison of the MGISEQ-2000 and Illumina HiSeq 4000 sequencing platforms for RNA sequencing. *Genomics & Informatics* **17(3)**:e32 DOI 10.5808/GI.2019.17.3.e32.

Jeon SA, Park JL, Park SJ, Kim JH, Goh SH, Han JY, Kim SY. 2021. Comparison between MGI and Illumina sequencing platforms for whole genome sequencing. *Genes Genomics* **43(7)**:713–724 DOI 10.1007/s13258-021-01096-x.

Kang W, Anslan S, Börner N, Schwarz A, Schmidt R, Künzel S, Rioual P, Echeverría-Galindo P, Vences M, Wang J, Schwalb A. 2021. Diatom metabarcoding and microscopic analyses from sediment samples at Lake Nam Co, Tibet: the effect of sample-size and bioinformatics on the identified communities. *Ecological Indicators* **121(6)**:107070 DOI 10.1016/j.ecolind.2020.107070.

Kelly M, Boonham N, Juggins S, Killie P, Mann D, Pass D, Sapp M, Sato S, Glover R. 2018. *A DNA based diatom metabarcoding approach for water framework directive classification of rivers*. Bristol: Environment Agency.

Kim HM, Jeon S, Chung O, Jun JH, Kim HS, Blazyte A, Lee HY, Yu Y, Cho YS, Bolser DM, Bhak J. 2021. Comparative analysis of 7 short-read sequencing platforms using the Korean reference genome: MGI and Illumina sequencing benchmark for whole-genome sequencing. *GigaScience* **10(3)**:giab014 DOI 10.1093/gigascience/giab014.

Korostin D, Kulemin N, Naumov V, Belova V, Kwon D, Gorbachev A. 2020. Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. *PLOS ONE* **15(3)**:e0230301 DOI 10.1371/journal.pone.0230301.

**Kueneman JG, Bletz MC, McKenzie VJ, Becker CG, Joseph MB, Abarca JG, Archer H, Arellano AL, Bataille A, Becker M, Belden LK, Crottini A, Geffers R, Haddad CFB, Harris RN, Holden WM, Hughey M, Jarek M, Kearns PJ, Kerby JL, Kielgast J, Kurabayashi A, Longo AV, Loudon A, Medina D, Nuñez JJ, Perl RGB, Pinto-Tomás A, Rabemananjara FCE, Rebollar EA, Rodríguez A, Rollins-Smith L, Stevenson R, Tebbe CC, Vargas Asensio G, Waldman B, Walke JB, Whitfield SM, Zamudio KR, Zúñiga Chaves I,

Woodhams DC, Vences M. 2019. Community richness of amphibian skin bacteria correlates with bioclimate at the global scale. *Nature Ecology & Evolution* **3(3)**:381–389 DOI 10.1038/s41559-019-0798-1.

Kumar KR, Cowley MJ, Davis RL. 2019. Next-generation sequencing and emerging technologies. In: *Seminars in thrombosis and hemostasis*, Thieme Medical Publishers, 661–673.

Leray M, Knowlton N. 2017. Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ* **5**:e3006.

Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ. 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology* **10(1)**:34 DOI 10.1186/1742-9994-10-34.

Li Q, Zhao X, Zhang W, Wang L, Wang J, Xu D, Mei Z, Liu Q, Du S, Li Z. 2019. Reliable multiplex sequencing with rare index mis-assignment on DNB-based NGS platform. *BMC Genomics* **20(1)**:1–13 DOI 10.1186/s12864-019-5569-5.

Liu K, Liu Y, Hu A, Wang F, Chen Y, Gu Z, Anslan S, Hou J. 2020. Different community assembly mechanisms underlie similar biogeography of bacteria and microeukaryotes in Tibetan lakes. *FEMS Microbiology Ecology* **96(6)**:fiaa071 DOI 10.1093/femsec/fiaa071.

Loit K, Adamson K, Bahram M, Puusepp R, Anslan S, Kiiker R, Drenkhan R, Tedersoo L. 2019. Relative performance of MinION (Oxford Nanopore Technologies) versus sequel (Pacific Biosciences) third-generation sequencing instruments in identification of agricultural and forest fungal pathogens. *Applied and Environmental Microbiology* **85(21)**:e01368-19 DOI 10.1128/AEM.01368-19.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* **17(1)**:10–12 DOI 10.14806/ej.17.1.200.

McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE* **8(4)**:e61217 DOI 10.1371/journal.pone.0061217.

Nilsson RH, Anslan S, Bahram M, Wurzbacher C, Baldrian P, Tedersoo L. 2019. Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nature Reviews Microbiology* **17(2)**:95–109 DOI 10.1038/s41579-018-0116-y.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara R, Simpson GL, Solymos P, Stevens M, Wagner H. 2015. R package 'vegan': community ecology package. *Available at https://cran.r-project.org, https://github.com/vegandevs/vegan*.

Oliveros J. 2018. Venny: an interactive tool for comparing lists with Venn's diagrams. *Available at https://bioinfogpcnbcsices/tools/venny/indexhtml*.

Pont D, Valentini A, Rocle M, Maire A, Delaigue O, Jean P, Dejean T. 2021. The future of fish-based ecological assessment of European rivers: from traditional EU water framework directive compliant methods to eDNA metabarcoding-based approaches. *Journal of Fish Biology* **98(2)**:354–366 DOI 10.1111/jfb.14176.

Porter TM, Hajibabaei M. 2018. Automated high throughput animal CO1 metabarcode classification. *Scientific Reports* **8**:1–10.

R-Core-Team. 2021. *R: a language and environment for statistical computing*. Vienna: The R Foundation for Statistical Computing. *Available at https://www.R-project.org/*.

Reitmeier S, Hitch TC, Treichel N, Fikas N, Hausmann B, Ramer-Tait AE, Neuhaus K, Berry D, Haller D, Lagkouvardos I. 2021. Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling. *ISME Communications* **1**:1–12.

Rocchi S, Valot B, Reboux G, Millon L. 2017. DNA metabarcoding to assess indoor fungal communities: electrostatic dust collectors and Illumina sequencing. *Journal of Microbiological Methods* **139(6)**:107–112 DOI 10.1016/j.mimet.2017.05.014.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4(17)**:e2584 DOI 10.7717/peerj.2584.

Schnell IB, Bohmann K, Gilbert MTP. 2015. Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources* **15(6)**:1289–1303 DOI 10.1111/1755-0998.12402.

Stan-Development-Team. 2021. Stan modeling language users guide and reference manual, v2.21. *Available at https://mc-stan.org*.

Sun X, Hu Y-H, Wang J, Fang C, Li J, Han M, Wei X, Zheng H, Luo X, Jia Y. 2021. Efficient and stable metabarcoding sequencing data using a DNBSEQ-G400 sequencer validated by comprehensive community analyses. *Gigabyte* **1(5667)**:1–15 DOI 10.46471/gigabyte.16.

Taberlet P, Bonin A, Coissac E, Zinger L. 2018. *Environmental DNA: for biodiversity research and monitoring*. Oxford: Oxford University Press.

Tedersoo L, Albertsen M, Anslan S, Callahan B. 2021. Perspectives and benefits of long reads and synthetic long reads in microbial ecology. *Applied and Environmental Microbiology* **87(17)**:e0062621 DOI 10.1128/AEM.00626-21.

Valentin V, Frédéric R, Isabelle D, Olivier M, Yorick R, Agnès B. 2019. Assessing pollution of aquatic environments with diatoms' DNA metabarcoding: experience and developments from France Water framework directive networks. *Metabarcoding and Metagenomics* **3**:e39646 DOI 10.3897/mbmg.3.39646.

Watts C, Dopheide A, Holdaway R, Davis C, Wood J, Thornburrow D, Dickie IA. 2019. DNA metabarcoding as a tool for invertebrate community monitoring: a case study comparison with conventional techniques. *Austral Entomology* **58(3)**:675–686 DOI 10.1111/aen.12384.

Wen C, Wu L, Qin Y, Van Nostrand JD, Ning D, Sun B, Xue K, Liu F, Deng Y, Liang Y, Zhou J. 2017. Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. *PLOS ONE* **12(4)**:e0176716 DOI 10.1371/journal.pone.0176716.

Wurzbacher C, Fuchs A, Attermeyer K, Frindte K, Grossart H-P, Hupfer M, Casper P, Monaghan MT. 2017. Shifts among Eukaryota, Bacteria, and Archaea define the vertical organization of a lake sediment. *Microbiome* **5**:41.

Zawierucha K, Porazinska DL, Ficetola GF, Ambrosini R, Baccolo G, Buda J, Ceballos JL, Devetter M, Dial R, Franzetti A, Fuglewicz U, Gielly L, Lokas E, Janko K, Jaromerska TN, Koscinski A, Kozlowska A, Ono M, Parnikoza I, Pittino F, Poniecka E, Sommers P, Schmidt SK, Shain D, Sikorska S, Uetake J, Takeuchi N. 2021. A hole in the nematosphere: tardigrades and rotifers dominate the cryoconite hole environment, whereas nematodes are missing. *Journal of Zoology* **313(1)**:18–36 DOI 10.1111/jzo.12832.