Bioinformatics for Biomonitoring: Species Detection and Diversity Estimates Across Next-Generation Sequencing Platforms

Isaac M.K. Eckert[*], Joanne E. Littlefair[*,1], Guang K. Zhang[*], Frédéric J.J. Chain[*,†], Teresa J. Crease[‡], Melania E. Cristescu[*]

[*]Department of Biology, McGill University, Montréal, QC, Canada
[†]Department of Biological Sciences, University of Massachusetts Lowell, Lowell, MA, United States
[‡]Department of Integrative Biology, University of Guelph, Guelph, ON, Canada
[1]Corresponding author: e-mail address: joanne.littlefair@mail.mcgill.ca

Contents

Advances in Ecological Research
ISSN 0065-2504
https://doi.org/10.1016/bs.aecr.2018.06.002

Abstract

As a fast-growing area of technology, sequencing platforms are updated frequently and this rapid technical revolution poses not only great advances but also challenges. To be effective, biomonitoring programmes need to deliver comparable results across research groups and time. Understanding the sources of bias in bioinformatics promotes reliable results that accurately reflect biodiversity. We assembled two mock communities of planktonic organisms to assess the accuracy of species recovery based on sequencing the 18S rRNA V4 region using two NGS platforms, Roche 454 (the platform of choice for early metabarcoding studies), and Illumina MiSeq (employed frequently in recent metabarcoding studies). Our findings suggest that the two platforms have comparable performance on metabarcoding datasets. When singletons (sequences represented by a single read) were excluded from analyses, Illumina MiSeq had a slightly better operational taxonomic unit (OTU) precision score than Roche 454 (calculated as the number of species detected divided by the number of OTUs generated) but only in one bioinformatics workflow (when paired reads were appended, not merged). Roche 454 performed slightly better than Illumina MiSeq in terms of species detection but only when simple mock communities with a single individual per species were analysed. When singleton sequences were included, both platforms detected more than 75% of species with a slightly higher detection achieved by Illumina MiSeq. The OTU clustering of both datasets resulted in a gross overestimation of species richness. This finding suggests that studies employing OTU clustering as a proxy for genetic diversity must carefully perform read processing, such as singleton exclusion, to avoid overestimates. Finally, this study provides insight into technical bioinformatic strategies that should accompany such transitions. In a field such as metabarcoding, where advances in sequencing technology constantly drive the discipline, ensuring the comparability of past and future technologies, and the derived ecological conclusions is important.

1. INTRODUCTION

Metabarcoding has the potential to become a powerful tool for rapid biodiversity assessment, describing long-term biodiversity trends, studying the ecology and evolution of natural communities, and developing new, rapid, and efficient techniques for biomonitoring (Cristescu, 2014; Littlefair and Clare, 2016). At the core of metabarcoding is next-generation sequencing (NGS) technology, which allows the sequencing of mixed, complex environmental samples (Pompanon et al., 2012; Taberlet et al., 2012). The main advantage of NGS over the traditional Sanger sequencing is that it provides greater sequencing depth, high-resolution analyses, and eliminates the need to generate single-individual libraries prior to sequencing. These advantages make NGS approaches attractive for biomonitoring studies. However, the high depth comes with the cost of reduced length of sequenced reads (Leigh et al., 2015) and a relatively high error rate (Goodwin et al., 2016). Metabarcoding approaches using NGS allow the parallel examination of multiple taxonomic groups through the use of universal barcodes (specified short fragments of DNA), without a great deal of a priori knowledge about the targeted organisms. Species assignments are often conducted by matching the retrieved barcodes against large open-access reference libraries of DNA sequences that are constantly being populated with reference barcodes (e.g., BOLD, an informatics workbench containing open-access COI barcode records; SILVA, an open-access database which curates small (e.g., 16S/18S) and large (e.g., 23S/28S) subunit ribosomal RNA sequences).

Within the scope of biomonitoring, metabarcoding approaches have the potential to save time and alleviate the problem of deficient taxonomic expertise in the identification of more obscure groups (Deiner et al., 2017); as such expertise is often distributed unevenly around the globe (Ji et al., 2013). It has also been shown that metabarcoding approaches can provide taxonomic information with increased resolution in relation to existing monitoring protocols. For example, NGS techniques provided higher taxonomic resolution at lower cost than morphological identification using Environment Canada's Canadian Aquatic Biomonitoring Network protocols (Gibson et al., 2015). This "big data" approach to biodiversity science allows us to focus on multiple taxonomic groups, rather than monitoring with indicator species, the use of which can be problematic if not carefully linked to ecosystem functioning and measures of true diversity (Moonen and Bàrberi, 2008). If combined with almost real-time DNA sequencers currently in development (e.g., MinION, GridION), metabarcoding has the potential to provide very fine scale temporal and spatial monitoring data from around the globe (Bohan et al., 2017). The power derived from greater amounts of data will allow us to monitor ecological networks and their properties, from which we can infer or model information about ecosystem structure and stability (Evans et al., 2016). Governments are starting to consider integrating molecular methods into existing monitoring programmes (Darling and Mahon, 2011; Kelly et al., 2014). For example, in 2014 the UK government approved the use of environmental DNA as an alternative to conventional surveys to monitor the great crested newt, whose habitats are protected from development by European and UK law. The European Union has begun the integration of DNA-based tools into European ecological monitoring programmes by developing a series of work packages known as DNAqua-NET (Leese et al., 2016; see chapter "Next generation biomonitoring of aquatic ecosystems" by Leese). However, to integrate NGS into existing management strategies, we need to evaluate the consistency of biodiversity estimates based on metabarcoding datasets as sequencing platforms and bioinformatic tools are replaced with ever-evolving technology.

While the tremendous amount of data produced by NGS is advantageous, the accompanying need for stringent and specific filtering of data is paramount, with appropriate recording of processing steps when it comes to producing repeatable and reliable results and confirming species detection (Clare et al., 2016). Additionally, bioinformatic pipelines associated with metabarcoding continue to change rapidly with evolving NGS technology. Rapid change can be problematic when such technologies become integrated into long-term biomonitoring programmes, where consistency of, and comparison between results is valued (Coissac et al., 2012). In particular, NGS platforms are constantly updated as technological innovations become available, providing more accurate and in-depth analyses (Glenn, 2011; Goodwin et al., 2016; Zhou et al., 2013). This emphasizes the need to set appropriate guidelines on the use of sequencing platforms and bioinformatics.

To date, the major platforms that have been used in metabarcoding studies (Table 1) are Roche 454, Illumina HiSeq, and MiSeq, and Ion Torrent (Heather and Chain, 2016). The Roche 454 sequencing platform produces continuous single reads by employing pyrosequencing. This involves the addition of a nucleotide base onto a growing nucleotide chain coupled with the enzymatic release of light, which is monitored by a camera and recorded

Table 1 A Summary of the Next-Generation Sequencing (NGS) Technologies Discussed in This Chapter, Their Output Characteristics, and Their Main Types of Errors

NGS Technology	Output: Read Length	Output: Number of Reads	Error Rate %	Main Error Type
Roche 454	Single reads: 400–700 bp	>1 million	1	Insertion/ deletion
Illumina MiSeq	Paired reads: 2 × 300 bp	25 million	>0.1%	Substitution
Illumina HiSeq	Paired reads: 2 × 150 bp	>300 million	≥0.1%	Substitution
Ion Proton 1	Single reads: 200 bp	60 million	1%	Insertion/ deletion

Modified from Glenn, T.C., 2011. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* 11 (5), 759–769.

as one of the four DNA nucleotides (Balzer et al., 2010; Leamon et al., 2003). Illumina MiSeq and HiSeq both use the same method of dye labelling and sequencing by synthesis. Fluorescently labelled bases release fluorescence when incorporated into a growing nucleotide chain, which is detected by a laser and recorded. Unlike Roche 454, Illumina platforms can also produce paired reads that represent the start and end of a target amplicon (Bentley, 2008). These paired reads can be merged to produce continuous sequences (contigs) if they overlap with one another (e.g., the fragment size is shorter than the combined paired read length). HiSeq was designed to provide much higher data output compared to MiSeq, and can produce a much greater number of raw reads, but sacrifices short run time and long read length to do so. Ion Torrent sequencing relies on the monitored release of hydrogen atoms from a growing strand of nucleotides (hydrogen atoms released into the reaction solution change the pH, causing a signal to be recorded by an ion-sensitive field-effect transmitter). This platform can produce continuous reads at a much lower cost compared to other NGS technologies (Rothberg et al., 2011).

Despite the importance of understanding sources of potential bias in the lab and bioinformatics steps within metabarcoding (Alberdi et al., 2017; Brown et al., 2015; Clare et al., 2016; Flynn et al., 2015), the taxonomic bias introduced through the use of different sequencing technologies has received relatively little attention, particularly within the scope of metazoan communities. Conclusions from studies using genomics and transcriptomics applications suggest a variety of platform-specific biases due to the differing technologies (Table 2). For example, Quince et al. (2009) found a higher error rates associated with Roche 454 when sequencing homopolymer regions. Although Illumina improves on this problem, it suffers from its own limitations. For example, reads of different quality can be produced depending on their location on the sequencing plate (Erlich et al., 2008). While Illumina reads are shorter than Roche reads, the ability to use 300bp paired reads to sequence both ends of a fragment gives the option of targeting the longer amplicons that can be analysed with Roche. Long continuous regions of the genome assembled from Illumina reads have been shown to be more accurate and complete when compared to Roche assemblies, due to the increased read depth (Luo et al., 2012). Despite these differences, Luo et al. (2012) found that Illumina and Roche are comparable in terms of the ability to assess the diversity of a microbial community from a pooled genetic sample. Furthermore, while Tremblay et al. (2015) identified a taxonomic bias associated with NGS platforms and the ability to detect

Table 2 A Summary of Comparative Studies of Sequencing Platforms

Publication	Platforms Investigated	Methods and Applications	Conclusions
Clooney et al. (2016)	Illumina MiSeq, Illumina HiSeq, Ion PGM	Microbiome analysis of stool samples for bacterial species detection	• HiSeq shotgun libraries had the greatest read depth • HiSeq shotgun sequences identified the highest number of species • MiSeq and Ion PGM provided a better basis for identifying functional genes due to longer read length
Mahé et al. (2015)	Illumina MiSeq and Roche 454	Comparison of quality and quantity of reads in an environmental diversity assessment	• MiSeq produced an order of magnitude more reads • More amplicons of different taxonomic identity (above the genus level) were found with MiSeq due to deeper sequencing capabilities • Roche 454 produced reads of better quality; therefore, more raw reads were retained after filtering
Tremblay et al. (2015)	Illumina MiSeq and Roche 454	Analysis of amplified V4, V6–V8, and V7–V8 reads from a microbial mock community, sequenced with both platforms	• Paired end MiSeq reads produced higher quality data and allowed for more aggressive quality control parameters, resulting in a higher retention rate of reads for further analysis • The impact of sequencing platform bias was relatively minor compared to the bias introduced by primer selection
Luo et al. (2012)	Illumina Genome Analyzer II and Roche 454	Compared base-call error, frameshift frequency and contig length between Illumina and Roche 454 sequencing data produced from a complex freshwater planktonic community	• Illumina yielded longer and more accurate contigs (fewer truncated genes due to frameshifts) • Roche 454 produced assemblies that contained a significantly higher proportion of frameshift errors compared to Illumina assemblies from the same genome • Both sequencing platforms were reliable for quantitatively assessing genetic diversity within natural communities

Li et al. (2014)	Illumina and Roche 454	Sequencing identical libraries composed of plasma from patients failing antiretroviral therapy for HIV	• Illumina data resulted in higher coverage as well as increased sensitivity for detecting HIV-1 minority variants • Illumina also produced fewer false-positive variant calls compared to Roche 454
Salipante et al. (2014)	Illumina MiSeq and Ion Torrent PGM	Comparison of sequencers for 16S rRNA-based bacterial community profiling in terms of differences in error rates, read truncation, and species detection	• Both platforms were comparable for species detection with only minor differences • Ion Torrent PGM exhibited much higher error rates as well as a pattern of premature sequence truncation • The suggested cause of premature sequence truncation by Ion PGM was the secondary structure of the sequences, as no identifiable primary sequence pattern was found to exist among the truncated sequences
Divoll et al. (2018)	Illumina MiSeq and Ion Torrent PGM	Compared the ability of both sequencers to resolve a list of prey species from bat faecal matter, using similar library preparation and identical analytical workflows	• 104 prey OTUs were detected by both platforms • 176 prey OTUs were detected by MiSeq only • 17 prey OTUs were detected by Ion Torrent only • Results suggest that Illumina MiSeq greatly outperformed Ion Torrent in terms of species resolution from a sample of community DNA

There is a paucity of eukaryotic metabarcoding studies that directly compare Illumina MiSeq and Roche 454 in terms of species detection, biomonitoring, and diversity estimates.

species within a mock community, the bias was relatively minor, compared to the taxonomic bias introduced through primer selection (Tremblay et al., 2015). The authors conclude that Illumina is advantageous due to the higher quality (lower insertion and deletion error rates) of data produced, resulting in a higher retention rate of reads for final analysis.

Early biomonitoring studies based on metabarcoding used Roche 454 technology (e.g., Geml et al., 2014; Hatzenbuhler et al., 2017; Lallias et al., 2015) due to its long read length and simplified bioinformatics, which did not require merging steps for forward and reverse reads. Later, Illumina MiSeq technologies were adopted due to the attraction of greater read depth (Evans et al., 2017; Hänfling et al., 2016). However, platform-specific differences exist in terms of read depth and error rate, as well as specific bioinformatics steps necessary to account for the inherent difference in read length. The recent transition from Roche to Illumina as the most commonly used NGS platform raises questions of comparability for long-term biomonitoring studies aiming to assess changes in community composition. If metabarcoding techniques are integrated into long-term monitoring schemes, it is important to examine how technological transitions will impact observational results obtained by examining experimental communities with defined species compositions. For example, increased read depth might influence the sensitivity for species detection in complex samples, and whether rare species and invasion fronts can be detected. Higher error rates can obscure taxonomic identities when interspecies divergence is low. These questions are not just important for the comparability of sequencing technologies (Mahé et al., 2015; Tremblay et al., 2015), but also for the field of metabarcoding as a whole. Standardization of a particular NGS platform for biomonitoring, while it has its advantages, could prevent researchers from using the most suitable or up-to-date platform for the focus of their study.

Once NGS libraries are sequenced, raw reads go through several bioinformatics quality control steps before species identity can be assigned with confidence. These include demultiplexing (assigning mixed sequences back to their original samples), read length filtering, quality control filtering, and clustering into operational taxonomic units (OTUs) (Floyd et al., 2002). OTUs can then be used to identify species based on alignment with a species reference sequence. Species abundance has been correlated to both sequenced read counts and OTU richness (Lim et al., 2016). Transforming the hundreds of thousands of raw reads produced in a sequencing run into a conclusive list of taxa is a daunting task, and also provides many opportunities for bias

(Schmidt et al., 2013). The analytical procedures need to be tailored to the study and can, if misapplied, result in an overestimation of biodiversity (Brown et al., 2015; Flynn et al., 2015). Understanding the possible introductions of bias at each step in a metabarcoding workflow is essential when it comes to generating reliable results including sample collection, molecular techniques, and bioinformatics tools (Alberdi et al., 2017; Clare et al., 2016; Flynn et al., 2015; Pawluczyk et al., 2015).

In this study, we compare two NGS technologies in the context of biomonitoring of zooplankton communities. As NGS technologies are steadily evolving, it is important to consider whether we can achieve consistent results over long-term studies that use molecular methods. One major recent transition involved the shift from the discontinued Roche 454 to Illumina MiSeq as the platform of choice for metabarcoding studies. Here, we use Roche 454 and Illumina MiSeq to sequence two mock zooplankton communities (artificial communities in which the composition of species is known) targeting the hypervariable V4 region of the 18S rRNA gene (18S-V4 region). We analyse the data using very similar bioinformatics pipelines for both datasets. We used workflows and pipelines originally tested by Flynn et al. (2015) with one additional step to deal with merging or appending paired reads from the MiSeq. This study compares and contrasts the number of reads, species detection results, and precision scores generated from both NGS platforms with the aim of qualifying the transition from Roche to Illumina sequencing for biomonitoring studies using metabarcoding.

2. MATERIALS AND METHODS

2.1 Mock Communities

We used two mock communities assembled by Brown et al. (2015). The simple community 1 was represented by a single individual per species and consisted of 56 species of zooplankton: 46 arthropods, 2 chordates, and 8 molluscs. The complex community 2 was represented by populations and consisted of a total of 76 individuals belonging to 14 species: 12 crustaceans and 2 molluscs (Table S1 in the online version at https://doi.org/10.1016/bs.aecr.2018.06.002). Mock communities were assembled to avoid congeneric species, with some exceptions (Balanus, Daphnia, Hyallela, and Gammarus). Individuals included in the mock communities were identified to species or genus level based on morphology and Sanger sequencing. Although eight individuals could only be identified to the family level, we nevertheless included them in the community since these individuals were taxonomically and genetically divergent from other community members. The individual specimens, preserved in ethanol, were washed with distilled water prior to assembly. To increase the efficiency of the DNA extraction involving a large number of individuals, both mock communities were assembled in four microcentrifuge tubes. Each tube, containing approximately equal numbers of individuals, was then centrifuged to remove any liquid that remained after the specimens were washed. Genomic DNA was extracted from each of the four tubes using DNeasy Blood and Tissue Kit (Qiagen, Venlo, Limburg, Netherlands) following the manufacturer's protocol. The DNA extractions were then used to prepare NGS libraries to be run on two separate sequencing platforms to profile the zooplankton mock community, the Roche 454 and the Illumina MiSeq.

2.2 Library Preparation for Roche 454

DNA amplification of the 18S-V4 region was performed, and the amplicons were sequenced using the Roche 454 (Brown et al., 2015). PCR was performed using universal 18S-V4 primers to generate a 400–700 bp fragment (F: AGGGCAAKYCTGGTGCCAGC, R: GRCGGTATCTRATCGYCTT) (Zhan et al., 2013). To reduce PCR amplification bias, eight reactions were performed on each independent DNA extraction, and equimolar aliquots of each replicate were pooled for sequencing. PCR reactions consisted of 100 ng of genomic DNA, 1 × PCR buffer, 2 mmol/L of Mg^{2+}, 0.2 mmol/L of dNTPs, 0.4 μmol/L of each primer, and 2 U of Taq DNA polymerase (Genscript, Piscataway, NJ, USA), in a final reaction volume of 25 μL.

PCR cycling parameters consisted of an initial denaturation step at 95°C for 5 min, followed by 25 amplification cycles of 95°C for 30 s, 50°C for 30 s, 72°C for 90 s, and a final elongation step at 72°C for 10 min. PCR products were cleaned using the solid-phase reversible immobilization paramagnetic bead method (ChargeSwitch, Invitrogen, Carlsbad, CA, USA). Quantity and quality were assessed using gel electrophoresis and the Quant-iT PicoGreen dsDNA assay kit (Invitrogen, Carlsbad, CA, USA). Cleaned PCR products were then pooled together in equimolar concentrations before pryosequencing at the 1/2 PicoTiter plate scale. Roche 454 flex adapters (A: CCATCTCATCCCTGCGTGTCTCCGACTCAG, B: CCTATCCCCTGTGTGCCTTGGCAGTCTCAG) were added to the 5′ end of the 18S-V4 amplicons. Sequencing was performed using 454 FLEX Adapter A on a GS-FLEX Titanium platform (454 Life Sciences, Branford, CT, USA) by Génome Québec (Montréal QC, Canada). The data generated include 483,986 reads for community 1 (accession SRX884895) and 204,922 reads for community 2 (accession SRX884904) with an average length of 518 bp and can be found in the Sequence Read Archive.

2.3 Library Preparation for Illumina MiSeq

Library preparation for the MiSeq was performed by first PCR amplifying the 18S-V4 region with gene-specific primers as in the Roche workflow. The Forward gene-specific primer was as follows: Adapter–Spacer[Gene-Specific Region]: 5′-TCGTCGGCAGCGTC-AGATGTGTATAAGAGACAG [AGGGCAAKYCTGGTGCCAGC]-3′. The Reverse gene-specific primer was as follows: Adapter–Spacer[Gene-Specific Region]: 5′-GTCTC GTGGGCTCGG-AGATGTGTATAAGAGACAG[GRCGGTATCTR ATCGYCTT]-3′. All the initial PCR products were visualized on a 1.5% agarose gel and submitted to the Genomics Facility at the University of Guelph for further processing. The initial PCR products were purified using AMPure beads and indexed by PCR amplification with primers containing the index sequences; Forward primers were (Flowcell Adapter–Index–Adapter): 5′-AATGATACGGCGACCACCGAGATCTACAC-INDEX-TCGTCGG CAGCGTC-3′ and Reverse primers were (Flowcell Adapter–Index–Adapter): 5′-CAAGCAGAAGACGGCATACGAGAT-INDEX-GTCT CGTGGGCTCGG-3′. An equal volume of the eight PCR amplicons from each of the four samples from each community was pooled before performing the index amplification. Two libraries were created for each community from the pooled templates for a total of four pooled libraries. The indexed libraries were again purified with AMPure beads, quantified, normalized, and pooled for sequencing using the paired-end 2 × 300 bp cartridge on the MiSeq (Illumina, Inc., San Diego, CA, USA). The data generated include 1,005,261 paired raw reads and can be found in the Sequence Read Archive, National Center for Biotechnology Information NCBI ID: SRR6848116 (Community 1) and SRR6848115 (Community 2).

2.4 Bioinformatics and Data Analysis

The data were processed in Unix using a series of data filtering and analysis packages (Table 3), and were compared to evaluate the species detection and performance of the Roche and MiSeq platforms. For all pipelines, taxonomic assignment was performed using BLAST alignments against a local reference database that was constructed to contain an 18S-V4 reference sequence for all members of the mock community, which were sufficiently

Table 3 The Major Steps of Each Bioinformatics Workflow for the Two Sequencing Platforms (Roche 454 or Illumina MiSeq) and the Main Workflows: With Singletons Included (S) or Excluded (N), and Merging (M) or Appending (AP) Steps for Illumina MiSeq

Bioinformatic Steps	**Illumina MiSeq**				**Roche 454**	
	N, M	**N, AP**	**S, M**	**S, AP**	**N**	**S**
Adapter removal and trimming (Trimmomatic)	✓	✓	✓	✓	✓	✓
Merging (FLASH)	✓	×	✓	×	×	×
Appending (FASTX)	×	✓	×	✓	×	×
Primer Removal (FASTX)	✓	✓	✓	✓	✓	✓
Dereplication (USEARCH)	✓	✓	✓	✓	✓	✓
Chimera removal (UCHIME)	✓	✓	✓	✓	✓	✓
Singleton exclusion (USEARCH)	✓	✓	×	×	✓	×
Abundance sorting	✓	✓	✓	✓	✓	✓
Clustering (UPARSE)	✓	✓	✓	✓	✓	✓
Taxonomic assignment (BLAST)	✓	✓	✓	✓	✓	✓

divergent such that each species could be individually identified. The local reference database was the same database used by Brown et al. (2015), assembled using reference sequences downloaded from the NCBI nucleotide database and the SILVA database (Quast et al., 2013) as well as sequences generated for closely related species using Sanger sequencing of the 18S-V4 region.

We received demultiplexed reads from both Génome Québec (Roche) and the University of Guelph (MiSeq). Reads from the two replicate libraries, for each mock community, were pooled prior to analysis. For both Roche and MiSeq pipelines, adapters were removed and reads were trimmed using Trimmomatic v0.32 (Bolger et al., 2014) based on Phred quality scores (a method of assessing the probability of an incorrect base call). Leading (5′) and trailing (3′) bases were trimmed if they had a quality score of less than 10, representing a 90% base accuracy threshold. In addition, bases were removed from a sequence if the average quality of a sliding window of 20 base pairs (bp) fell beneath an average quality score of 10. Finally, sequences that were shorter than 100 bp were discarded.

Assembly of MiSeq paired end reads into contigs occurred using two distinct methods: merging reads (M) and blunt-end appending reads (AP).

The overlapping and merging of paired end reads were conducted in Flash v1.2.7 (Magoč and Salzberg, 2011). Reads that did not merge were discarded, based on the expectation that the target amplicon should be smaller than the combined read pairs (~570 bp), which would mean the read pairs would overlap. Since the merging step can introduce a bias against species with long 18S-V4 regions if read pairs do not overlap, the impact of merging MiSeq reads was investigated through analysis of blunt-end appended paired reads to recover species with long V4 regions that are filtered out due to insufficient read overlap during the merging step. Appended reads were created using FASTX-Toolkit 0.0.13.2 (http://hannonlab.cshl.edu/fastx_toolkit/) by reverse complementing the reverse read, so both paired reads were in the same orientation (5′-3′), and then joining them together in the middle without overlap.

The 18S-V4 amplicon primers were removed in both the Roche and MiSeq pipelines using the FASTX clipper. Reads that did not meet a minimum length of 200 bp were discarded prior to dereplication (collapsed into a set of unique sequences) and chimeras were filtered out using UCHIME (Table 4, Edgar et al., 2011). Singletons were either retained or discarded using USEARCH (Edgar, 2010). Datasets were analysed with OTU clustering using UPARSE with a 3% divergence threshold implemented in USEARCH (Edgar, 2013) following Brown et al. (2015). Direct taxonomic assessment of dereplicated sequences without OTU clustering was performed (Table 4) by aligning dereplicated sequences independently against reference sequences in the local reference database using BLAST. OTUs were taxonomically assigned using the same method against the same local reference database.

2.5 Assigning Taxonomy to OTUs and Dereplicated Sequences

The taxonomic classification of each dereplicated sequence and each OTU was based on the best BLAST hit against a local reference database used by Brown et al. (2015), which was defined as a hit with at least 90% identity and an alignment of at least 300 nucleotides with a reference sequence in the assembled local reference database. A threshold of 90% was chosen to accommodate congeneric reference sequences that represented species only identified to the family level, and a minimum alignment of 300 bp was chosen based on the methods used in Brown et al. (2015). We ensured that species-specific reference sequences for species belonging to these genera were included in our local database to allow identification

Table 4 A Summary of the Reads/Sequences Retained at Each Quality Filtering Step for the Mock Communities (1 and 2) Sequenced With the Roche 454 or Illumina MiSeq Platforms

Workflow	**Illumina MiSeq-1**		**Illumina MiSeq-2**		**Roche 454-1**	**Roche 454-2**
Merged (M) or Appended (AP)	**M**	**AP**	**M**	**AP**	**N/A**	**N/A**
Raw reads	543,024	543,024	462,237	462,237	483,986	204,922
Trimmed reads	348,216	538,871	298,597	458,611	151,196	46,764
Dereplicated sequences	212,701	523,656	179,562	438,918	58,941	19,518
Sequences after chimera filtering	199,303	517,828	173,718	436,371	58,717	19,483
Percentage of reads that are chimeras	4.6	1.1	2.2	0.6	0.2	0.1
Sequences after removing singletons	12,959	3171	8973	2956	7321	2411
Percentage singletons	56	97	57	95	34	37
Total filtered reads	332,074	533,003	292,055	456,063	150,942	46,722
Nonsingleton filtered reads	144,980	18,334	126,976	22,636	99,501	29,620

M = merged paired-end reads. AP = appended paired-end reads.

at the species level. We recorded the number of reads that matched a species, as well as the number of OTUs that returned each species (Table S1 in the online version at https://doi.org/10.1016/bs.aecr.2018.06.002). Finally, we compared the results from the two sequencing platforms in terms of numbers of OTUs produced, number of species detected, and assigned precision scores. Species detection was calculated as a percentage (number of species detected by a workflow, divided by number of species in the mock community). OTU precision was calculated as the number of species detected divided by the number of OTUs generated. An OTU precision score of 1.0 indicates that each OTU correctly corresponded 1:1 to the species included in the mock community with no extra or missing OTUs, while a low precision score indicates the presence of many additional OTUs. We examined read depth, OTU diversity estimates, singleton OTUs, and species detection for the Roche data and for each treatment of the MiSeq reads (merged or appended).

3. RESULTS

3.1 Sequence and Read Depth

The MiSeq runs produced greater numbers of retained reads and dereplicated sequences compared to the Roche runs, despite similar numbers of raw reads (Table 4). After filtering, dereplication, and excluding singletons, MiSeq produced 12,959 and 8973 (communities 1 and 2, respectively) successfully merged sequences compared to 7321 and 2411 filtered Roche sequences. However, fewer reads (3171 and 2956) were retained in the pipeline that involved appended MiSeq reads. Including singletons in the analysis produced a larger proportion of new sequences from MiSeq data, compared to Roche. Furthermore, final tallying of filtered merged reads shows that the MiSeq produced a greater proportion of singletons (56% and 57%) compared to Roche (34% and 37%) (Fig. 1). After assigning taxonomy, MiSeq produced an average read depth per species of 5000 and 15,101 (communities 1 and 2) excluding singletons, and 7532 and 25,013 including singletons. Roche produced an average read depth per species of 2584 and 2506 excluding singletons, and 3408 and 3678 including singletons (Table S1 in the online version at https://doi.org/10.1016/bs.aecr.2018.06.002). Excluding singletons, 144,961 (27% of raw reads, 99% of filtered reads) and 120,815 (26%, 95%) MiSeq reads successfully matched a reference sequence in our local BLAST database, while Roche only produced 98,203 (20%, 99%) and 25,065 (12%, 85%) reads that matched a reference sequence (Table 5). After including singletons, these numbers increased to 331,392 (61%, 99%) and 275,147 (59%, 94%) for MiSeq, and 146,566 (30%, 97%) and 40,456 (20%, 87%) for Roche 454.

3.2 OTU Clustering and the Effect of Singletons

Similar numbers of OTUs were identified from the Roche and MiSeq data when singletons were excluded (merged MiSeq: 53 and 26; Roche: 57 and 18; Table 5). However, the number of OTUs created during the clustering of dereplicated reads was significantly increased by the inclusion of singletons, which produced 18,120 and 16,269 OTUs (communities 1 and 2, respectively) from the merged MiSeq data and 459 and 154 OTUs from the Roche data. When singletons were included, the number of OTUs created from the MiSeq reads increased by three orders of magnitude, compared to an order of magnitude increase for Roche. With both platforms, the exclusion of singletons produced OTU estimates that are comparable to the number of species that were present in the mock community, while the inclusion of singletons created a large discrepancy between number of OTUs and species in the mock community. Upon taxonomical assignment, 83% and 42% (communities 1 and 2) of merged MiSeq OTUs successfully matched a species in the mock community when singletons were excluded, compared to 84% and 78% of Roche OTUs. When singletons were included, 99% and 94% of merged MiSeq OTUs matched a species in the mock community, compared to only 63% and 79% of Roche OTUs. The vast majority of singleton OTUs generated from MiSeq data matched a species in the mock community.

3.3 Species Detection

Only species included in the mock community were represented by a reference sequence in the constructed local BLAST database. The number of species detected in this study refers to the number of species present in the mock communities and detected by a best BLAST hit, excluding parasites, bacteria, or contaminants that were not included in the local database.

The large variability in OTU clustering of the Roche and MiSeq data was not reflected in the species detected mainly because we focused on false negatives (species present but not detected) and not false positives (species absent but detected). Detection scores were calculated as the number of species recovered divided by the number of species in the mock community (Table 5). Eleven of the 56 species in community 1 and 1 of the 14 species

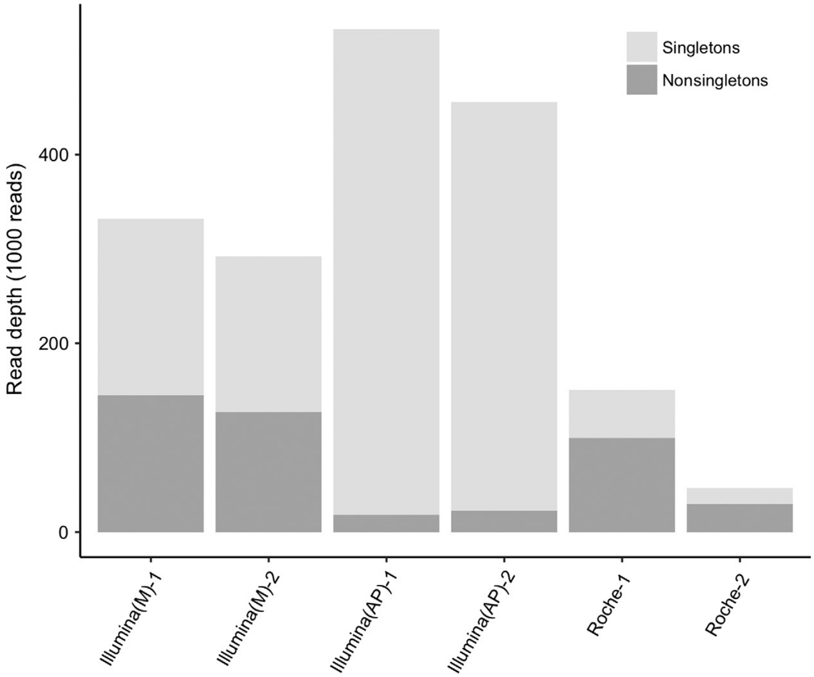

Fig. 1 The depth (in thousands) of filtered reads from Illumina MiSeq and Roche 454 sequencing of mock zooplankton communities. There is increased singleton generation in the MiSeq platform, especially in workflows that append reads (AP), instead of merging (M). The *x-axis labels* refer to the mock communities; simple community 1 with 56 species, each represented by a single individual and the complex community 2 with 14 species, represented by populations of individuals (Table S1 in the online version at https://doi.org/10.1016/bs.aecr.2018.06.002).

in community 2 were never detected using either platform, with or without including singletons. This could have been due to poor DNA extraction, but the undetected species from community 1 were spread across the four subcommunities (Table S1 in the online version at https://doi.org/10.1016/bs.aecr.2018.06.002) suggesting that the tissue for these species may have been of low quality.

When singletons were excluded, fewer species were detected in the merged MiSeq data (29 species; detection score [DS] = 0.518 in community 1, and 8 species; DS = 0.571 in community 2) compared to Roche (38 species; DS = 0.679 in communities 1 and 10 species; DS = 0.714 in community 2; Fig. 2A). The inclusion of singletons increased the number of species detected in both communities with the merged MiSeq data (44 species; DS = 0.786 for communities 1 and 11 species; DS = 0.786 for community 2)

Table 5 Comparison of Results From the Two NGS Platforms (Roche 454 or Illumina MiSeq)

Sequencing Platform	**Illumina MiSeq-1**				**Illumina MiSeq-2**				**Roche 454-1**		**Roche 454-2**	
S or N	**N**	**S**	**N**	**S**	**N**	**S**	**N**	**S**	**N**	**S**	**N**	**S**
Merged or Appended	**M**	**M**	**AP**	**AP**	**M**	**M**	**AP**	**AP**	**N/A**	**N/A**	**N/A**	**N/A**
Reads that matched a BLAST hit	144,961	331,392	18,315	502,435	120,815	275,147	22,240	385,658	98,203	146,566	25,065	40,456
Number of species detected	29	44	30	44	8	11	10	13	38	43	10	11
Number of OTUs	53	18,120	35	258,914	26	16,269	15	220,207	57	459	18	154
Percentage of singleton OTUs	0	85	0	91	0	88	0	93	0	48	0	44
Number of OTUs that matched a species in the mock community	44	17,924	34	243,640	11	15,283	12	161,706	48	287	14	122
Percentage of raw reads that match a reference sequence	27	61	3.4	93	26	60	4.8	83	20	30	12	20
Percentage of filtered reads that match a reference sequence	99	99	99	94	95	94	98	85	99	97	85	87
Percentage of OTUs that matched a species in the mock community	83	99	97	94	42	94	80	73	84	63	78	79
Detection score	0.518	0.786	0.536	0.786	0.571	0.786	0.714	0.929	0.679	0.769	0.714	0.786
OTU precision	0.547	0.003	0.857	0	0.308	0.001	0.667	0	0.667	0.094	0.556	0.071

Singletons were either included in (S) or excluded from (N) the analyses. Illumina MiSeq reads were either merged traditionally (M) or appended (AP) to form contigs. Precision score refers to the number of species recovered divided by the number of operational taxonomic units (OTU) generated. Detection score was calculated as the number of species recovered divided by the number of species in the mock community. Precision was calculated as the number of species recovered divided by the number of OTUs generated.

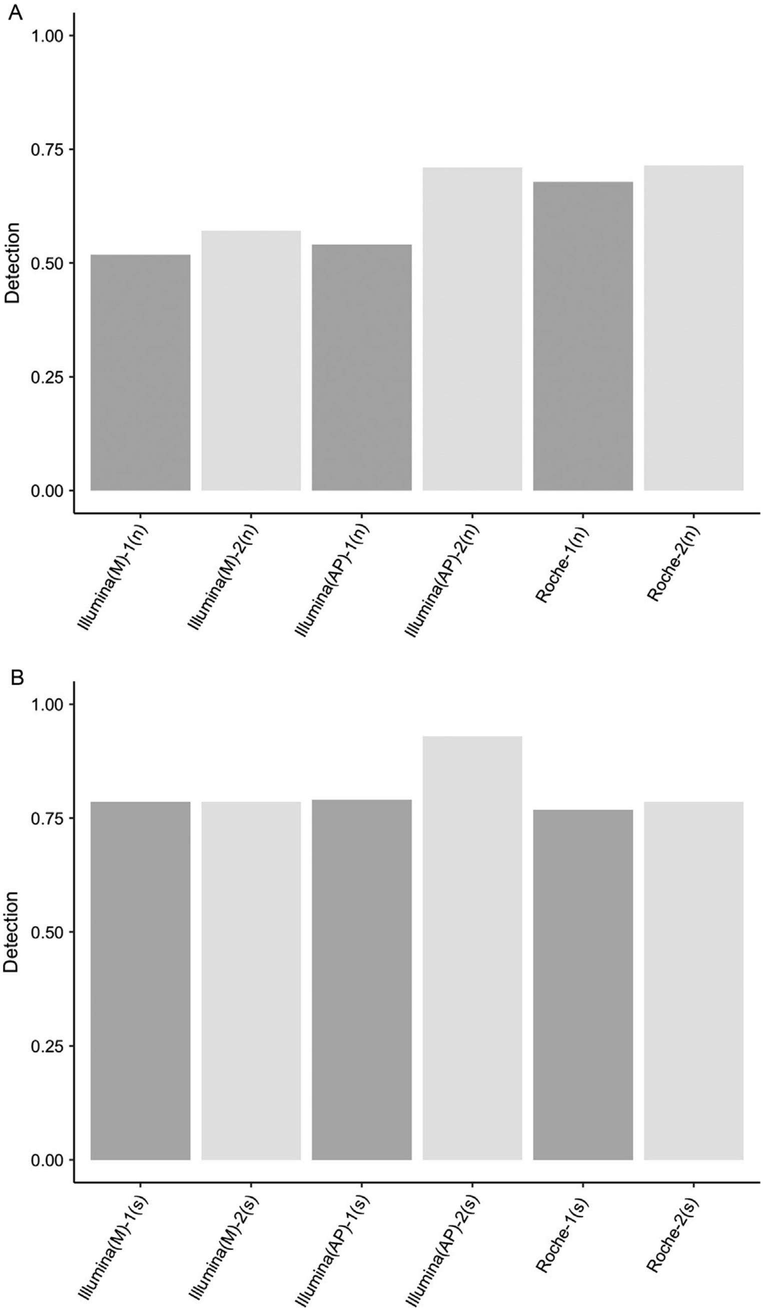

Fig. 2 The detection of species in mock communities by different next-generation sequencing platforms and bioinformatic workflows. Values were calculated by dividing the number of species recovered by the number of species in the mock community. (A) Workflows that exclude singletons. (B) Workflows that include singletons. Species detection varied with choice of platform and workflow between the simple community 1 and the complex community 2. The *x-axis labels* refer to the choice of platform, the workflow (M = merged MiSeq reads, AP = appended MiSeq reads), the mock community (1 in *dark grey* or 2 in *light grey*) and the inclusion(s) or exclusion(n) of singletons.

and Roche (43 species; DS = 0.769 for communities 1 and 11 species; DS = 0.786 for community 2; Fig. 2B). We recovered one species (*Hyalella* clade 8) with the MiSeq data that was never detected using the Roche data, and one species (*Hyperoche mediterranea*) that was only recovered with Roche data. With the inclusion of singletons, we recovered 15 and 3 (communities 1 and 2) additional species with merged MiSeq data compared to 5 and 1 additional species with Roche data.

3.4 OTU Precision

The increased ability to detect species with singleton inclusion was accompanied by a large decrease in precision (Fig. 3, Table 5), which was calculated as the number of species recovered divided by the total number of OTUs generated (including OTUs that did not match a species in the mock community). The precision value represents the inverse of the average number of OTUs created per species, where a smaller precision score means more OTUs per species, or more OTUs that do not match a species in the mock community. Merged MiSeq data produced consistently lower OTU precision scores (0.547 and 0.308, communities 1 and 2 without singletons; 0.002 and 0.001 with singletons) compared to Roche data (0.667 and 0.556 without singletons; 0.094 and 0.071 with singletons). After excluding singletons from the analysis, Roche produced higher precision scores than the merged MiSeq data, but lower scores than the appended MiSeq data (Fig. 3). The appended MiSeq data including singletons produced the lowest precision scores. Low precision scores are due to multiple OTUs being generated for a single species (Table S1 in the online version at https://doi.org/10.1016/bs.aecr.2018.06.002) and are prevalent even when a stringent filtering/clustering approach (exclusion of singletons and a 3% divergence threshold for OTU clustering) is used.

3.5 Impact of Merging and Appending MiSeq Reads on Species Detection and OTU Estimates

Merging MiSeq reads requires overlapping sequences, which will generally not occur with long amplicon sequences due to read length limitations imposed by the sequencing platform. Even so, some species with long 18S-V4 regions were detected with the merged data (Table S1 in the online version at https://doi.org/10.1016/bs.aecr.2018.06.002). However, MiSeq read pairs can be appended to one another to create potentially discontinuous, but informative sequences. The appending technique joins paired

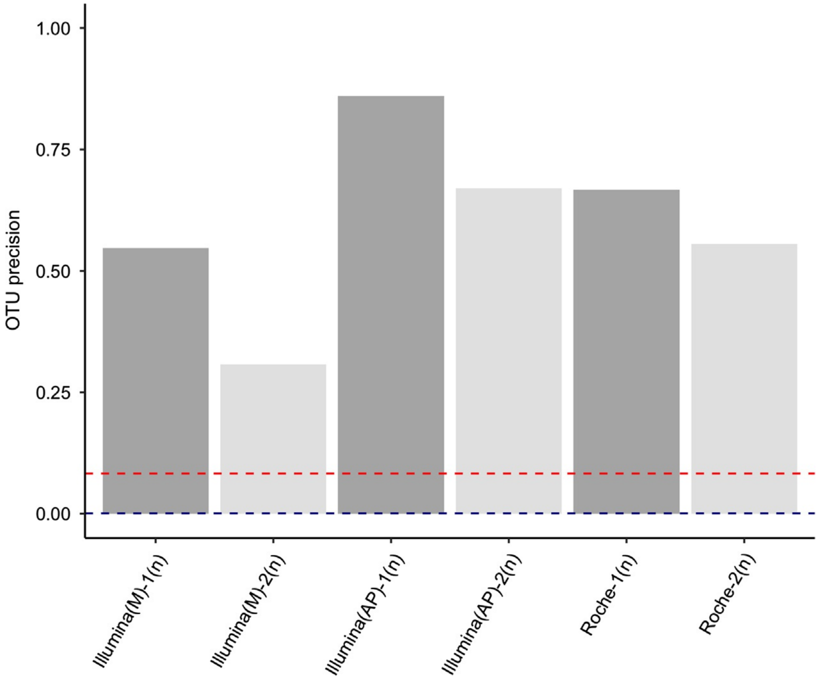

Fig. 3 Precision scores estimated from different bioinformatic workflows analysing next-generation sequence data from two mock zooplankton communities. Singletons were excluded from the analysis. Precision is defined as the number of species recovered divided by the number of operational taxonomic units generated. The *x-axis labels* refer to the choice of platform, the bioinformatic workflow (M = merged MiSeq reads; AP = appended MiSeq reads) and the mock community (simple community 1 in *dark grey* and complex community 2 in *light grey*). The *red and blue lines* indicate the average precision including singletons for Roche and MiSeq workflows, respectively.

reads together without overlap to generate a contig. In the appending process, reads that would overlap and merge contain a repeated sequence in the middle where the overlap would have occurred, and the length of the overlap can vary. In addition, paired-end reads that do not overlap may not always end at the same nucleotide position. As a result, there is substantial variation among appended sequences and they do not dereplicate into as few sequences as merged sequences. For this reason, appending pipelines generate a much higher number of singletons (Table 4).

There was little difference in the number of species detected with merged and appended reads, but more species were detected when singletons were included. When singletons were excluded, 30 species (DS = 0.536) were detected in community 1 with appended reads compared to 29 species

(DS = 0.518) with merged reads (Table 5). Ten species (DS = 0.714) were detected in community 2 with appended reads compared to 8 species (DS = 0.571) with merged reads. When singletons were included, 44 species (DS = 0.786) were detected in community 1 with both merged and appended reads. Thirteen species (DS = 0.929) were detected in community 2 with appended reads compared to 11 species (DS = 0.786) with merged reads.

Clustering of appended reads into OTUs resulted in the generation of very high numbers of OTUs when singletons were included in the analysis, even more so than merged reads. Whereas appended reads produced 35 and 15 OTUs (communities 1 and 2) when singletons were excluded, 258,914 and 220,207 OTUs were produced when singletons were included (Table 5). Owing to the variation generated by the appending process, very low precision scores (1.69×10^{-4} and 5.90×10^{-5} for communities 1 and 2) were obtained when singletons were included compared to exclusion of singletons (0.857 and 0.667).

4. DISCUSSION

The Illumina MiSeq and the Roche 454 NGS platforms were used to sequence two mock communities, one simple community containing 56 species represented by a single individual per species and a more complex community containing 14 species represented by individuals or populations. The MiSeq runs produced more filtered reads compared to the Roche runs, and a larger percentage of singleton reads, from a similar amount of raw data. When singletons were excluded from the analysis, the MiSeq and Roche datasets generated similar numbers of OTUs. Moreover, the two platforms generated comparable species detection and OTU precision for both mock communities, although the species detection rate for community 1 was slightly higher with the Roche data. When singletons were included, one more species in community 1 and two more species in community 2 were detected with the appended MiSeq reads compared to the Roche data. However, when singletons were included, Roche produced much higher number of OTUs than species included in the communities (459 for community 1 and 154 for community 2). Even higher numbers of OTUs were generated with the MiSeq data when singletons were included resulting in very low precision scores. Even so, the majority of OTUs matched a species in the mock community (99% and 94% for merged and appended reads in community 1, and 94% and 73% in community 2).

With one exception, these values are considerably higher than those obtained for Roche OTUs (63% and 79% for communities 1 and 2).

Merging vs appending MiSeq paired reads had a small impact on species detection but a large impact on OTU precision as the precision scores for appended reads were orders of magnitude lower than those for merged reads. This is a consequence of the variation generated by the appending process in which forward and reverse reads are joined together even if they overlap (which is required for merging), and regardless of their length, as long as it meets the length threshold. This results in highly variable contigs due to the loss of quality control that occurs during merging because reads that do not meet an overlap threshold (a combination of length of overlap and sequence similarity of overlapped region) are discarded in merged workflows. Other bioinformatics studies have confirmed that merging Illumina reads helps decrease the numbers of spurious OTUs, as it improves predictions of the overlapping sequence when forward and reverse reads are aligned (Edgar and Flyvbjerg, 2015). While merging reads did substantially decrease the number of OTUs in our study, it also slightly decreased species detection, compared to appending reads.

4.1 Read Depth and Singletons

Despite a greater read depth, the MiSeq data did not provide an increase in species detection ability compared to the Roche. The inclusion of singletons, which can represent rare authentic sequences or sequencing artefacts, did aid species detection with both platforms. However, it also caused a large decrease in precision. Singletons provided detection of 15 and 3 (communities 1 and 2) additional species with merged MiSeq data, and 5 and 1 additional species with Roche 454 data that were not detected in workflows excluding singletons (Table 5). Most singletons detected a species already recovered when included in taxonomic analysis. The relationship between inclusion of singletons, an increase in number of species detected, and a decrease in precision is conserved across sequencing platforms, although it is more pronounced with the MiSeq data. This suggests that the MiSeq platform is more susceptible to singleton generation (it generated a higher percentage of singletons), especially when appending reads. Most studies do exclude singleton sequences (Abad et al., 2016; Chain et al., 2016; although see Gibson et al., 2015). However, studies focused on the detection of rare species such as aquatic invaders often include singletons in their analysis.

4.2 OTU Clustering

The number of OTUs generated was extremely variable between sequencing platforms and between the workflows involving the inclusion and exclusion of singletons. However, the number remained relatively consistent within each platform in terms of order of magnitude, and for results from communities 1 and 2. A greater number of OTUs was consistently produced from the MiSeq data compared to Roche, which resulted in lower precision scores. Many of the additional MiSeq OTUs matched the same species in the mock community. For example, when reads from community 1 (excluding singletons) were clustered into OTUs and taxonomically assigned, four OTUs matching *Artemia* spp. were identified in the MiSeq data, while only one was identified in the Roche data.

The generation of more than one OTU per species is due to increased sequence variation produced by either genuine intragenomic variation (genetic variation within the genome of a single individual/species) or sequencing artefacts. OTU inflation needs to be taken into consideration when designing a metabarcoding workflow and when generating conclusions. This study found that with singletons included, MiSeq produced a much higher number of OTUs compared to Roche. Flynn et al. (2015) reported very high OTU estimates with Roche data and suggested that terminal gaps are partially responsible. These gaps occur when the single read falls short of the reverse primer, which can contribute to divergence during OTU clustering. Analysis of different gap treatments using clustering algorithms that excluded and included terminal gaps supported this hypothesis (Flynn et al., 2015). However, merged and appended MiSeq reads do not contain terminal gaps. With paired read sequencing, the target amplicon is sequenced from both the 5′ and 3′ ends and the reads are merged or appended to produce a full contig. Therefore, it is likely that internal gaps generated during the merging and appending process are the source of OTU overestimation with MiSeq data. Studies that rely on OTU-based diversity estimates should consider excluding singletons to avoid gross overestimation.

When singletons were excluded from analysis, both platforms produced similar OTU estimates. However, some species were still represented by more than one OTU. For example, *Mytilus edulis* in community 1 was represented by five OTUs in the MiSeq data, but was only represented by one OTU in the Roche data. Conversely, *Leptodora kindtii* was represented by four OTUs in the Roche data for community 1, but only represented by one OTU in the MiSeq data. These results suggest that the sequencing platform seems to introduce some taxonomical bias with respect to inflating OTU diversity. One possible explanation for this phenomenon is the genetic variation between different copies of the marker sequence within a single genome. The 18S-V4 region is known for its hypervariability in terms of both nucleotide sequence and length (Wuyts et al., 2000, 2001). Moreover, the occurrence of variation at both the intraspecific and intra-individual levels has been well documented (Crease and Taylor, 1998; Hancock and Vogler, 2000). It has been suggested that slippage replication is an important mechanism causing length expansion in hypervariable regions (Crease and Taylor, 1998; Hancock and Vogler, 2000), and the maintenance of stem-loop structures in these regions may be a function of the mutational process itself, regardless of functional constraint to maintain function (Hancock and Vogler, 2000). Nevertheless, it is clear that at least some of the variation in the V4 sequences is of biological origin and must be taken into account when estimating OTU diversity.

4.3 Experimental Design

The construction of a mock community allows creation of a local reference BLAST database, which removes the need for OTU clustering. OTU clustering can reduce the time it takes to generate results when using large databases like NCBI or SILVA, where searching each dereplicated sequence against millions of reference sequences requires substantial computing time. In this study, the filtered dereplicated reads were aligned individually, using BLAST, against a local reference database and tallied after returning a hit. While the use of a local reference database is efficient and effective when studying mock communities of known composition, it is not applicable for metabarcoding projects, for example when there is no a priori knowledge of community composition.

The direct comparability of results from the MiSeq and Roche platforms depends on the similarity of the workflows used to analyse the sequence reads. Comparability of results between platforms is of high importance when considering the use of novel technologies in long-term studies. Complete contigs can be generated from MiSeq paired-end reads that are usually merged together when the amplicon is short enough that paired reads are expected to overlap. On the other hand, Roche data are always single ended. Although the use of appended MiSeq reads detected a slightly larger number of species compared to merged MiSeq reads, there are several reasons why these processing steps may not be suitable for some biodiversity studies.

When singletons were included, using appended reads overestimated the number of OTUs by over three orders of magnitude compared to the MiSeq and the Roche 454 estimates. Thus, the confidence of researchers in the species detection results generated from appended Illumina reads can also be questioned, especially if a local reference database cannot be used as in biodiversity surveys.

Excluding singletons from the workflow improves precision of appended reads to the point where appended workflows produce better precision scores than merged ones. However, these workflows sacrifice a considerable amount of read depth because there is considerable variation around the point where the paired reads are joined due to length variation among reads, and the fact that overlapping reads are joined instead of overlapped. This means that many appended sequences are unique and fail to cluster into OTUs. Therefore, the vast majority of appended reads exit the pipeline as singletons. For example, MiSeq data for community 1 contained 543,024 raw paired end reads. After appending and filtering, those raw reads produced 514,657 singleton sequences and only 3171 nonsingleton sequences compared to 12,959 nonsingleton merged sequences. Thus, merging allowed the retention of four times as many nonsingleton sequences compared to appending. While use of appending techniques as an alternative to merging may provide slightly higher species detection, researchers must compensate for the increased read diversity and decreased OTU precision that accompanies appended MiSeq workflows involving singletons, as well as the decreased sequence retention in workflows that exclude singletons.

Although the 18S-V4 region is commonly used as a phylogenetic marker for eukaryotes, there are several shortcomings of this barcode as a marker for metabarcoding that should be recognized. Ribosomal sequences are present in multiple copies, providing the opportunity for variation within the genome of a single individual (Bik et al., 2012). Decelle et al. (2014) assessed the intragenomic variability of the V4 region between different nuclei in two species of radiolarian and detected up to five OTUs in a single individual. However, the authors concluded that the analytical procedure used to obtain the results had a larger impact than the contribution of intragenomic variation to the presence of multiple OTUs per species. Even so, variation in length between gene copies is particularly common in expansion segments such as the V4 region (James et al., 2009), and such length variation can introduce bias against species with long amplicons during the MiSeq merging process. Indeed, the V4 region is longer than 570 bp in 12 (80%) of the 15 species in community 1 that were not detected with merged MiSeq data when singletons were excluded, but were detected with singleton inclusion. Thus, use of the 18S-V4 region as a genetic barcode should be taken into consideration when designing future studies because alternatives, such as the mitochondrial CO1 gene, exist. The transition from Roche to Illumina as the accepted NGS platform for metabarcoding adds importance to this idea by providing increased opportunity for overestimation of biodiversity.

4.4 Implications for Biomonitoring

The application of metabarcoding procedures to biodiversity, community ecology, and invasive species projects will depend on many sampling decisions, including study design, sample capture, and molecular and bioinformatic methods. Such decisions will be made in conjunction with the aims of each individual study. Increasingly, molecular methods will be integrated into ongoing biomonitoring campaigns, or may be used to start new biodiversity surveys (Leese et al., 2016). Methodological consistency is key when designing studies which range over temporal and spatial scales. Here, we have shown that changes in sequencing platform will produce different outcomes for species detection and OTU diversity estimates. It is likely that sequencing platforms and molecular methods will continue to evolve, probably at a greater pace than traditional methods, given their recent and rapid development. Therefore, it is likely that over a decades-long monitoring campaign, several sequencing platforms may be used.

Some studies aim to generate diversity estimates only, and prefer to assign OTUs or sequences to higher taxonomic ranks such as Order or Family. This may be because the primer pair in use cannot discriminate between taxa at lower taxonomic ranks, or because species-level reference sequences are not available, for example when work is conducted on tropical invertebrates (Salinas-Ramos et al., 2015). OTUs without species-level assignment can still be included in estimates of alpha and beta diversity, patterns of distribution and seasonality, and simple ecological analyses such as niche overlap (Clare et al., 2016). In this study, we demonstrated that the two sequencing platforms, Roche 454 and Illumina MiSeq produced comparable OTU values when singletons where excluded. Conversely, when singletons were included, vastly inflated OTU numbers were generated from the Illumina MiSeq data; sometimes several orders of magnitude greater than estimates based on Roche 454 data, as well as the true number of species included in the mock community. We therefore do not recommend comparisons of data produced by these platforms when singletons are included.

Many researchers choose to remove singletons from analysis (e.g., Port et al., 2016), but we detected some species only when including them. It is often emphasized that the inclusion of singletons could facilitate the detection of very rare species or invasion fronts (Brown et al., 2015; Jousset et al., 2017; Leray and Knowlton, 2017). Even so, Scott et al. (2018) recommend removal of singletons to reduce false-positive errors at the cost of slightly reduced sensitivity. The authors found that retaining singletons can reduce false negatives (species not detected) only when clustering is used prior to species assignment.

Many biomonitoring studies are focused on species detection as a starting point before exploring community richness and diversity. There may also be particular species of interest, for which detection as well as temporal and spatial distribution is important. While singleton inclusion might initially seem advantageous, researchers may not be able to use a local reference database to assign species identity in a biomonitoring study, but instead BLAST their results against a universal database such as BOLD or NCBI. It is likely that many of these additional OTUs might match to closely related species, or even OTUs of different taxa such as parasites or fungi, that while not intentionally included in the mock community, are inevitably present, and will likely be amplified by any barcoding study that uses universal primers.

5. CONCLUSIONS

Overall, the two NGS platforms used in this study generated comparable species detection and OTU precision scores. However, we detected unexpectedly high variation when using appended MiSeq reads, and when including singletons in the analysis. While the inclusion of singletons can aid in species detection, reducing false negatives, as demonstrated by both platforms, the generation of spurious OTU clusters that accompanies their inclusion can increase false positives and needs to be taken into account. The results from this study support the trends identified by Flynn et al. (2015) in their analysis of Roche data, as well as the impact of singletons identified in other studies (Clare et al., 2016). Specific analysis of the trade-off between accurate biodiversity estimation (singleton exclusion) and increased detection ability (singleton inclusion) should be considered when performing metabarcoding analyses. Furthermore, while the use of NGS technologies for metabarcoding shows great promise for the study of community composition, species detection, and biomonitoring; this study illustrates the need for tailored and cautious analytical procedures. While these results illustrate how the transition from the Roche to an Illumina platform is accompanied by increases in technological ability, and greater genomic resolution, Illumina sequencing is not without limitations since the short reads generated require additional bioinformatics manipulations that impact species detection. Future sequencing technologies and analytical tools may enable researchers to overcome the shortcomings of Illumina sequencing, and make further progress in the application and reliability of NGS platforms to the field of metabarcoding.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Jullien M. Flynn and Emily A. Brown for their advice with analysis and for preparing libraries for sequencing. This study was supported by NSERC Discovery grants to T.J.C. and M.E.C.

REFERENCES

Abad, D., et al., 2016. Is metabarcoding suitable for estuarine plankton monitoring? A comparative study with microscopy. Mar. Biol. 163 (7), 1–13.

Alberdi, A., Aizpurua, O., Bohmann, K., 2017. Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods Ecol. Evol. 9 (1), 134–147.

Balzer, S., et al., 2010. Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim. Bioinformatics 26, 420–425.

Bentley, D.R., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59.

Bik, H.M., et al., 2012. Sequencing our way towards understanding global eukaryotic biodiversity. Trends Ecol. Evol. 27 (4), 233–243.

Bohan, D.A., et al., 2017. Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. Trends Ecol. Evol. 32 (7), 477–487.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (15), 2114–2120.

Brown, E.A., et al., 2015. Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? Ecol. Evol. 5 (11), 2234–2251.

Chain, F.J.J., et al., 2016. Metabarcoding reveals strong spatial structure and temporal turnover of zooplankton communities among marine and freshwater ports. Divers. Distrib. 22 (5), 493–504.

Clare, E.L., et al., 2016. The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data. Genome 59 (11), 981–990.

Clooney, A.G., et al., 2016. Comparing apples and oranges?: next generation sequencing and its impact on microbiome analysis. PLoS One 11 (2), e0148028. Available at: https://doi.org/10.1371/journal.pone.0148028.

Coissac, E., Riaz, T., Puillandre, N., 2012. Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol. Ecol. 21, 1834–1847.

Crease, T.J., Taylor, D.J., 1998. The origin and evolution of variable-region helices in V4 and V7 of the small-subunit ribosomal RNA of branchiopod crustaceans. Mol. Biol. Evol. 15 (11), 1430–1446.

Cristescu, M.E., 2014. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends Ecol. Evol. 29 (10), 566–571. Available at: http://linkinghub.elsevier.com/retrieve/pii/S016953471400175X.

Darling, J.A., Mahon, A.R., 2011. From molecules to management: adopting DNA-based methods for monitoring biological invasions in aquatic environments. Environ. Res. 111, 978–988.

Decelle, J., et al., 2014. Intracellular diversity of the V4 and V9 regions of the 18S rRNA in marine protists (radiolarians) assessed by high-throughput sequencing. PLoS One 9 (8), e104297.

Deiner, K., et al., 2017. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. Mol. Ecol. 26 (21), 5872–5895.

Divoll, T.J., et al., 2018. Disparities in second-generation DNA metabarcoding results exposed with accessible and repeatable workflows. Mol. Ecol. Resour. 18 (3), 1–12. Available at: http://doi.wiley.com/10.1111/1755-0998.12770.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26 (19), 2460–2461.

Edgar, R.C., 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat. Methods 10 (10), 996–998.

Edgar, R.C., Flyvbjerg, H., 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 31 (21), 3476–3482.

Edgar, R.C., et al., 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27 (16), 2194–2200.

Erlich, Y., et al., 2008. Alta-cyclic: a self-optimizing base caller for next-generation sequencing. Nat. Methods 5 (8), 679–682.

Evans, D.M., et al., 2016. Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems. Funct. Ecol. 30 (12), 1904–1916.

Evans, N.T., et al., 2017. Fish community assessment with eDNA metabarcoding: effects of sampling design and bioinformatic filtering. Can. J. Fish. Aquat. Sci. 74 (9), 1362–1374.

Floyd, R., et al., 2002. Molecular barcodes for soil nematode identification. Mol. Ecol. 11 (4), 839–850.

Flynn, J.M., et al., 2015. Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. Ecol. Evol. 5 (11), 2252–2266.

Geml, J., et al., 2014. The contribution of DNA metabarcoding to fungal conservation: diversity assessment, habitat partitioning and mapping red-listed fungi in protected coastal *Salix repens* communities in the Netherlands. PLoS One 9 (6), e99852.

Gibson, J.F., et al., 2015. Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. PLoS One 10 (10), e0138432. Available at: https://doi.org/10.1371/journal.pone.0138432.

Glenn, T.C., 2011. Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. 11 (5), 759–769.

Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17 (6), 333–351.

Hancock, J.M., Vogler, A.P., 2000. How slippage-derived sequences are incorporated into rRNA variable-region secondary structure: implications for phylogeny reconstruction. Mol. Phylogenet. Evol. 14 (3), 366–374.

Hänfling, B., et al., 2016. Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. Mol. Ecol. 25, 3101–3119.

Hatzenbuhler, C., et al., 2017. Sensitivity and accuracy of high-throughput metabarcoding methods for early detection of invasive fish species. Sci. Rep. 7, 46393. Available at: https://doi.org/10.1038/srep46393.

Heather, J.M., Chain, B., 2016. The sequence of sequencers: the history of sequencing DNA. Genomics 107, 1–8.

James, S.A., et al., 2009. Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces cerevisiae* as revealed by whole-genome resequencing. Genome Res. 19, 626–635.

Ji, Y., et al., 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecol. Lett. 16 (10), 1245–1257. Available at: http://doi.wiley.com/10.1111/ele.12162.

Jousset, A., et al., 2017. Where less may be more: how the rare biosphere pulls ecosystems strings. ISME J. 11 (4), 853–862. Available at: http://www.nature.com/doifinder/10.1038/ismej.2016.174.

Kelly, R.P., Port, J.A., Yamahara, K.M., Martone, R.G., Lowell, N., et al., 2014. Harnessing DNA to improve environmental management. Science 344, 1455–1456.

Lallias, D., et al., 2015. Environmental metabarcoding reveals heterogeneous drivers of microbial eukaryote diversity in contrasting estuarine ecosystems. ISME J. 9 (5), 1208–1221. Available at: https://doi.org/10.1038/ismej.2014.213.

Leamon, J.H., et al., 2003. A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain. Electrophoresis 24, 3769–3777.

Leese, F., et al., 2016. DNAqua-Net: developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. Res. Ideas Outcomes 2, e11321. https://doi.org/10.3897/rio.2.e11321.

Leigh, M.B., Taylor, L., Neufeld, J.D., 2015. Clone libraries of ribosomal RNA gene sequences for characterization of microbial communities. In: McGenity, T.J., Timmis, K.N., Nogales, B. (Eds.), Hydrocarbon and Lipid Microbiology Protocols. Springer, pp. 127–154.

Leray, M., Knowlton, N., 2017. Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. PeerJ 5, e3006. Available at: https://peerj.com/articles/3006.

Li, J.Z., et al., 2014. Comparison of Illumina and 454 deep sequencing in participants failing raltegravir-based antiretroviral therapy. PLoS One 9 (3), e90485.

Lim, N.K.M., et al., 2016. Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and communities. R. Soc. Open Sci. 3, 160635.

Littlefair, J.E., Clare, E.L., 2016. Barcoding the food chain: sanger to high-throughput sequencing. Genome 59 (11), 946–958.

Luo, C., et al., 2012. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One 7 (2), e30087.

Magoč, T., Salzberg, S.L., 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27 (21), 2957–2963.

Mahé, F., et al., 2015. Comparing high-throughput platforms for sequencing the V4 region of SSU-rDNA in environmental microbial eukaryotic diversity surveys. J. Eukaryot. Microbiol. 62, 338–345.

Moonen, A.C., Bàrberi, P., 2008. Functional biodiversity: an agroecosystem approach. Agric. Ecosyst. Environ. 127, 7–21.

Pawluczyk, M., et al., 2015. Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. Anal. Bioanal. Chem. 407 (7), 1841–1848.

Pompanon, F., et al., 2012. Who is eating what: diet assessment using next generation sequencing. Mol. Ecol. 21 (8), 1931–1950.

Port, J.A., et al., 2016. Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. Mol. Ecol. 25 (2), 527–541.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596.

Quince, C., Lanzen, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., Read, L.F., Sloan, WT., 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. Nat. Methods 6 (9), 639–641.

Rothberg, J.M., et al., 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348–352.

Salinas-Ramos, V.B., et al., 2015. Dietary overlap and seasonality in three species of mormoopid bats from a tropical dry forest. Mol. Ecol. 24 (20), 5296–5307.

Salipante, S.J., et al., 2014. Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Appl. Environ. Microbiol. 80 (24), 7583–7591.

Schmidt, P.-A., et al., 2013. Illumina metabarcoding of a soil fungal community. Soil Biol. Biochem. 65, 128–132. Available at: https://doi.org/10.1016/j.soilbio.2013.05.014.

Scott, R., Zhan, A., Brown, E.A., Chain, F.J.J., Cristescu, M.E., Gras, R., MacIsaac, H.J., 2018. Optimization and performance testing of a sequence processing pipeline applied to detection of nonindigenous species. Evol. Appl. 11 (6), 891–905.

Taberlet, P.P., et al., 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. 21 (8), 2045–2050.

Tremblay, J., et al., 2015. Primer and platform effects on 16S rRNA tag sequencing. Front. Microbiol. 6, 771.

Wuyts, J., et al., 2000. Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA. Nucleic Acids Res. 28 (23), 4698–4708.

Wuyts, J., Van De Peer, Y., De Wachter, R., 2001. Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA. Nucleic Acids Res. 29 (24), 5017–5028.

Zhan, A., et al., 2013. High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. Methods Ecol. Evol. 4 (6), 558–565.

Zhou, X., et al., 2013. Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. GigaScience 2 (4), 1–12.