

SCIENTIFIC REPORTS

OPEN

Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater

G. A. C. Singer¹, N. A. Fahner¹, J. G. Barnes¹, A. McCarthy¹ & M. Hajibabaei^{1,2}

The characterization of biodiversity is a crucial element of ecological investigations as well as environmental assessment and monitoring activities. Increasingly, amplicon-based environmental DNA metabarcoding (alternatively, marker gene metagenomics) is used for such studies given its ability to provide biodiversity data from various groups of organisms simply from analysis of bulk environmental samples such as water, soil or sediments. The Illumina MiSeq is currently the most popular tool for carrying out this work, but we set out to determine whether typical studies were reading enough DNA to detect rare organisms (i.e., those that may be of greatest interest such as endangered or invasive species) present in the environment. We collected sea water samples along two transects in Conception Bay, Newfoundland and analyzed them on the MiSeq with a sequencing depth of 100,000 reads per sample (exceeding the 60,000 per sample that is typical of similar studies). We then analyzed these same samples on Illumina's newest high-capacity platform, the NovaSeq, at a depth of 7 million reads per sample. Not surprisingly, the NovaSeq detected many more taxa than the MiSeq thanks to its much greater sequencing depth. However, contrary to our expectations this pattern was true even in depth-for-depth comparisons. In other words, the NovaSeq can detect more DNA sequence diversity within samples than the MiSeq, even at the exact same sequencing depth. Even when samples were reanalyzed on the MiSeq with a sequencing depth of 1 million reads each, the MiSeq's ability to detect new sequences plateaued while the NovaSeq continued to detect new sequence variants. These results have important biological implications. The NovaSeq found 40% more metazoan families in this environment than the MiSeq, including some of interest such as marine mammals and bony fish so the real-world implications of these findings are significant. These results are most likely associated to the advances incorporated in the NovaSeq, especially a patterned flow cell, which prevents similar sequences that are neighbours on the flow cell (common in metabarcoding studies) from being erroneously merged into single spots by the sequencing instrument. This study sets the stage for incorporating eDNA metabarcoding in comprehensive analysis of oceanic samples in a wide range of ecological and environmental investigations.

The inventorying and monitoring of biological diversity is a fundamental component of ecological and environmental studies. Additionally, characterizing biodiversity is part of the environmental impact assessments and ongoing environmental monitoring that are required by industry operating in environmentally sensitive locations¹. Stakeholders are increasingly becoming more concerned about environmental stewardship, and this applies equally to the terrestrial², freshwater³, and marine environments⁴ and covers all major taxonomic groups. A recent United Nations conference (UN Biodiversity Conference, Egypt, November 2018), highlighted the need for monitoring and protecting biodiversity with the key message of “investing in biodiversity for people and planet”. Despite the extreme importance of these efforts, the technology for carrying out biodiversity assessments has remained static for decades, relying heavily on observational data and capturing whole organisms from their environment for morphological analysis. Unfortunately, these procedures are error-prone, time-consuming,

¹Centre for Environmental Genomics Applications, eDNAtec Inc., St. John's, NL, Canada. ²Centre for Biodiversity Genomics & Department of Integrative Biology, University of Guelph, Guelph, ON, Canada. Correspondence and requests for materials should be addressed to M.H. (email: hajibabaei@gmail.com)

Reference	Publication type	Substrate	Type of study	Sequencing platform	Average raw reads per amplicon per sample
31	Pre-print	Leech gut contents	Targeted (vertebrates)	Illumina MiSeq	Unknown
32	Peer-reviewed	Herbivore feces	General biodiversity	IonTorrent PGM	4,156 (post-filtering)
33	Peer-reviewed	Soil	Targeted (<i>Phytophthora</i>)	Roche GS Junior	Unknown
34	Peer-reviewed	Freshwater ecosystem	Targeted: planktonic protists	Illumina MiSeq	150,000*
35	Peer-reviewed	Turbid freshwater	Targeted (vertebrates)	Illumina MiSeq	7,181 (post-filtering)
36	Peer-reviewed	Sea water	General biodiversity	Roche GS-FLX	21,192* (post-filtering)
37	Pre-print	Brackish water	General biodiversity	Illumina MiSeq	200,185 (post-filtering)
38	Peer-reviewed	High-salinity lake	General biodiversity	Illumina HiSeq 2500	124,779 for bacteria ~37,000 for archaea and eukaryota (post-filtering)
39	Pre-print	Wheat and oilseed rape residues	General biodiversity	Illumina MiSeq	80,000* (raw)
40	Peer-reviewed	Scat	Targeted (truffle fungi)	Illumina MiSeq	<60,000*
41	Peer-reviewed	Marine sediment	Targeted (microbial)	Illumina MiSeq	35,663
42	Pre-print	Mock community	Targeted (Symbiodiniaceae)	Illumina MiSeq	6,309* (raw)
43	Peer-reviewed	Plant leaves	Targeted (fungi)	Illumina MiSeq	112,931*
44	Peer-reviewed	Freshwater	General diversity	IonTorrent PGM	3,978 (post-filtering)
45	Peer-reviewed	Shrimp stomach contents	General biodiversity	Illumina MiSeq	350,000*
46	Peer-reviewed	Soil	General biodiversity	Illumina MiSeq	61,256 (raw)
47	Peer-reviewed	Bat guano	General biodiversity	Illumina MiSeq	106,858
48	Peer-reviewed	Bee nest chambers	General biodiversity	Illumina MiSeq	4,736 (post-filtering)
49	Peer-reviewed	Sea water	Targeted (dinoflagellates)	Roche GS Junior	252,086* (raw)
50	Peer-reviewed	Soil	General biodiversity	Illumina MiSeq and HiSeq	880,000* (raw)

Table 1. A survey of recently published metabarcoding studies shows that the Illumina MiSeq is the most commonly used instrument to analyze these data, and that sequencing depth per sample varies widely but has a median of approximately 60,000 reads. Cases where the sequencing depth was not directly reported and had to be inferred indirectly are indicated with an asterisk.

expensive, and tend to ignore small but ecologically important flora and fauna simply because they are difficult to identify visually^{2,5}.

Over the past decade, increasing attention has been paid to the analysis of environmental DNA (eDNA)—a combination of DNA from whole cellular material or that is shed from organisms as they move through their environment. The existence of large reference databases (especially the common “DNA barcode” marker, cytochrome oxidase c subunit 1, or COI^{6,7}) with the power of modern DNA sequencing instruments, enables environmental metabarcoding—the identification of many individual species from a simple water or sediment sample. Environmental metabarcoding is much faster than conventional techniques, is less labour-intensive, does not rely on the expertise of taxonomists, and produces orders of magnitude more information⁸.

Many side-by-side comparisons have been made between traditional morphological assessments and eDNA-based assessments. In all cases, eDNA is capable of detecting far more taxa overall. However, many of these studies also find that some organisms detected by traditional methods in the environment fail to be detected through metabarcoding^{1,4,9–13}. There are a number of potential reasons for this discrepancy: the use of “universal” primers that don’t amplify some taxa as well as others⁵; employing markers that have biased representation in reference databases¹⁴; or an inadequate depth of sequencing to detect eDNA that is in low abundance. These factors are especially important when eDNA analyses are performed to track specific target organisms that might be present in low abundance in complex settings such as the oceans (e.g., endangered species or invasive species).

The goal of the present study was to investigate the influence of sequencing depth and more advanced workflow, including a patterned flowcell, offered by illumina’s NovaSeq platform on the ability to detect biological diversity present in a sample. To carry out this work, we analyzed samples on an Illumina MiSeq instrument at a sequencing depth that is typical of similar studies (Table 1), then analyzed those same PCR products on an Illumina NovaSeq 6000—the most advanced HTS instrument available today—with a sequencing depth approximately 700 times greater than that of the MiSeq. Not surprisingly, the NovaSeq detected many more taxa than the MiSeq: specifically, with one marker the NovaSeq detected 200% more metazoan families than the MiSeq. Contrary to our expectations, the NovaSeq still outperformed the MiSeq even when we subsampled the data to make depth-for-depth comparisons, suggesting that the NovaSeq has superior qualities beyond its much greater sequencing capacity.

Results

The Illumina MiSeq is currently the most popular metabarcoding platform. Twenty of the most-recently indexed papers in Google Scholar featuring the “metabarcoding” keyword were obtained in early November 2018 to perform a mini-metanalysis of the instrument most such studies are favouriting at the

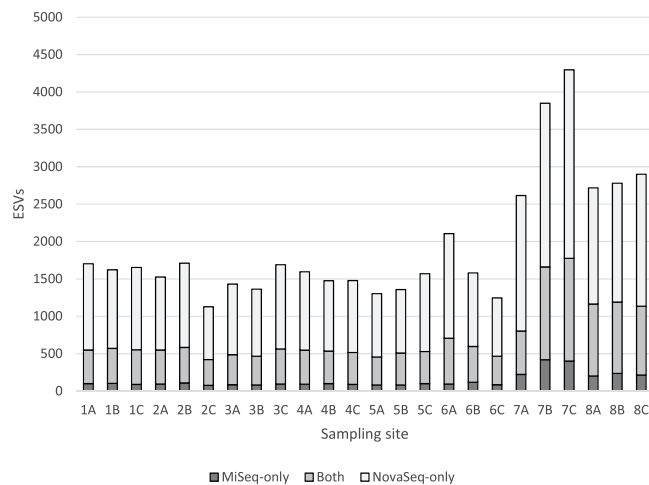


Figure 1. For each biological replicate (A–C) within each sampling site (1 to 8), the NovaSeq (light bars) was able to find a greater number of ESVs than the MiSeq (dark bars) even when the NovaSeq data are subsampled to match the sequencing depth of the MiSeq. Sites 7 and 8 are very close to shore, and site 7 in particular is near a wastewater outlet.

moment, as well as the depth of sequencing per sample that is generally employed. As shown in Table 1, the Illumina MiSeq is by far the instrument of choice presently, having been used by 14 (70%) of these studies. Sampling depth was not always reported clearly but was inferred where possible. Among these studies there was an extremely wide variance in sequencing depth, ranging from less than 10,000 reads per sample to nearly 900,000. However, the median was 60,000 (with a median absolute deviation of 55,000).

The NovaSeq finds more ESVs per sample than the MiSeq, even at the same sequencing depth.

Based on our literature survey (Table 1), we decided to analyze our own samples on the MiSeq with a targeted sequencing depth of 100,000 reads per amplicon per sample—approximately 50% greater than the median sequencing depth of those studies. Two amplicons were analyzed, FishE and F230, both located with the standard barcode region of the mitochondrial gene cytochrome oxidase C subunit 1 (COI). Post-filtering, the mean depth per sample was 118,290 reads for the FishE marker and 84,500 for the F230 marker. We then analyzed these same PCR products on the Illumina NovaSeq 6000 at much greater depth, averaging 7 million reads per amplicon per sample. The resulting reads were processed using the DADA2 pipeline as described in the Methods. Perhaps not surprisingly given the ~700x greater sequencing depth, the NovaSeq was able to find more exact sequence variants (ESVs) in each sample than the MiSeq. To our surprise, however, even after rarefying the NovaSeq data to match the sequencing depth of the MiSeq, we still found greater diversity (i.e., more ESVs) within the NovaSeq data for the FishE (Fig. 1) and F230 (Supplementary Fig. 1) amplicons. Moreover, while there was substantial overlap between the ESVs found between the two platforms, the MiSeq had very few ESVs unique to itself while the NovaSeq found many ESVs that the MiSeq missed. We highlight that the exact same PCR products were used for both instruments, so these results cannot be the consequence of stochastic PCR biases. The two sites with higher diversity—7 and 8—are located close to shore and site 7 in particular is close to a wastewater outlet (see Methods).

This trend is even more pronounced when plotted as an accumulation curve. When we combined all samples and then performed subsampling experiments to generate accumulation curves, the ability of the NovaSeq to detect new ESVs becomes even more stark: at each simulated sequencing depth, the NovaSeq detects greater biological diversity (i.e., ESVs) than the MiSeq (Fig. 2 for the FishE amplicon; see Supplementary Fig. 2 for the F230 amplicon). Curiously, while greater depth seems to reveal increasing numbers of ESVs on the NovaSeq (even beyond 2.5 million reads/sample), it is not clear that greater depth adds any new information for the MiSeq: the number of ESVs detected appears to level off at approximately 5,000. This strongly suggests that the NovaSeq outperforms the MiSeq in a depth-independent manner.

This trend is not an artefact of the DADA2 error-correcting model. DADA2 generates ESVs by applying an error correction model to raw FASTQ files, attempting to fix errors that were introduced through PCR or sequencing¹⁵. However, while the MiSeq reports base call qualities using pseudo-continuous Phred scores that can range from 0–40, the NovaSeq’s FASTQ files bin qualities into just four levels¹⁶. We therefore suspected that the phenomenon we were observing might be an artefact of the DADA2 program. Specifically, we hypothesized that the algorithm might be under-correcting errors in the NovaSeq data leading to a spurious increase in the number of ESVs. For this reason we repeated our analysis with simple OTU clustering at a 97% similarity threshold (described in greater detail in the Methods). OTU clustering applies no error correction model at all and is simply based on sequence similarity measures, and should therefore have the same performance on NovaSeq data as it does on MiSeq data. To our surprise, when accumulation curves were generated to compare the two instruments depth-for-depth, the NovaSeq once again outperformed the MiSeq (Fig. 3).

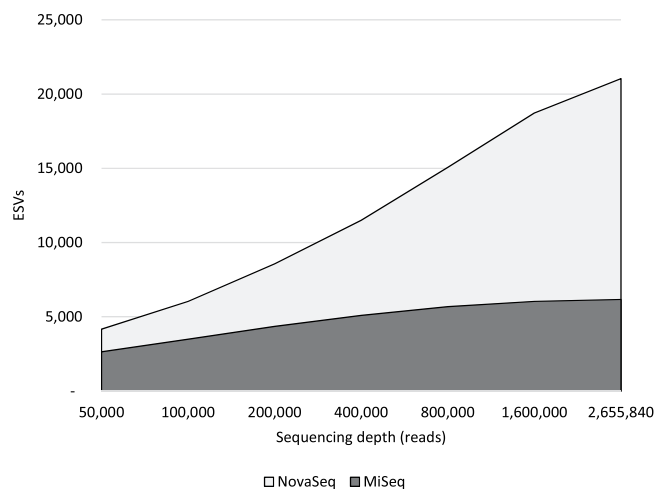


Figure 2. Accumulation curves generated by subsampling the MiSeq (dark curve) and NovaSeq data (light curve) for the FishE amplicon shows that depth-for-depth the NovaSeq detects greater biological diversity than the MiSeq.

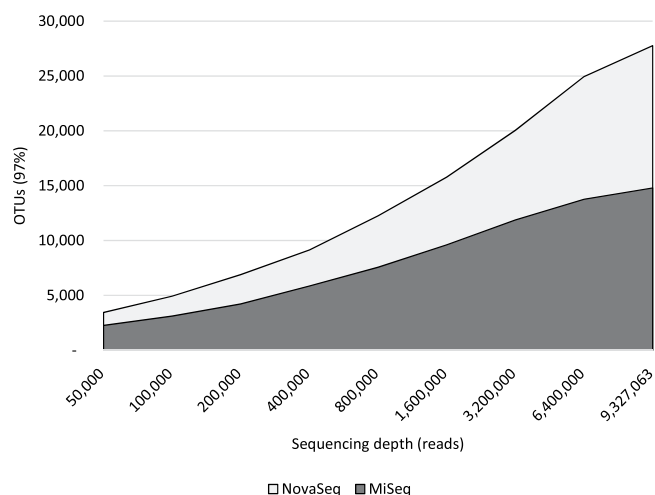


Figure 3. Accumulation curves of MiSeq (dark curve) and NovaSeq data (light curve) based on OTU clustering with a 97% identity threshold. At each sequencing depth, the NovaSeq finds more OTUs than the MiSeq, similar to the ESV data.

We note that Fig. 2 and Fig. 3 look quite different from each other in two ways: (1) the number of OTUs detected is far greater than the number of ESVs; and (2) while the number of new ESVs levels out for the MiSeq in Fig. 2, the trajectory continues upward for the OTUs in Fig. 3. This is due to the very different methodologies employed to generate ESVs versus OTUs. Because OTU clustering makes no attempt to model and correct for PCR and sequencing errors, the raw number of OTUs is expected to be much greater than the number of ESVs detected—many OTUs are simply the product of the accumulation of errors. By similar reasoning, both the MiSeq and NovaSeq OTU curves continue to rise with greater sequencing depth because additional sequencing errors will be encountered with that greater depth.

Greater sequencing depth on the MiSeq cannot achieve the level of diversity detected on the NovaSeq. The MiSeq's accumulation curve in Fig. 2 suggests that additional sequencing depth would not increase the number of ESVs detected, but to thoroughly test this point we re-ran three samples (sites 1, 3, and 6, each with three biological replicates for a total of 9 samples) on the MiSeq at much greater sequencing depth—approximately 1 million reads per amplicon per sample—and then compared these data to the NovaSeq data. As illustrated for the FishE amplicon in Fig. 4, adding this greater sequencing depth to the MiSeq only marginally improves its detection of diversity from the samples. Conversely, the NovaSeq continues to detect new ESVs with greater sequencing depth. Note that the total number of ESVs detected is lower than that of Fig. 2, but this is because of the smaller number of samples analyzed in this experiment (three sites versus eight). Again, we repeated this experiment with the F230 amplicon and found the same trend (see Supplementary Fig. 3).

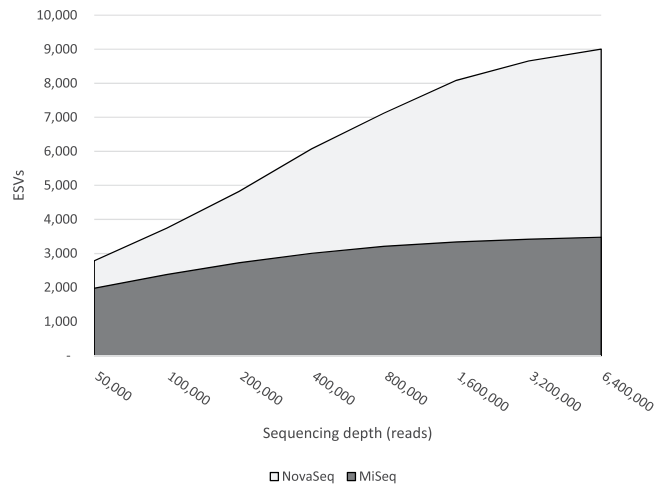


Figure 4. Re-running a subset of our samples on the MiSeq (dark curve) with much greater sequencing depth only added a very small number of new ESVs. Conversely, greater sequencing depth continues to add new ESVs to the NovaSeq data (light curve).

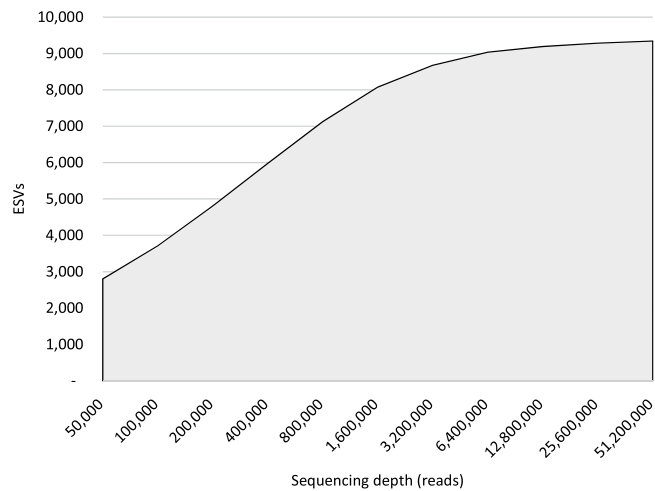


Figure 5. Accumulation curve of pooled NovaSeq data. The NovaSeq does eventually reach a plateau where no new ESVs are detected, albeit at an extreme sequencing depth of ~10 million reads.

Figure 4 shows that even beyond 5 million reads the NovaSeq was still finding new ESVs with no sign of plateauing. As before, we suspected this might be the result of the DADA2 algorithm under-correcting sequencing errors in the NovaSeq data. To examine this possibility, we ran the accumulation curve out to its maximum and found that the curve does indeed hit a plateau of just over 9,000 ESVs at a sequencing depth of approximately 10 million reads (Fig. 5). This result indicates that the pattern observed for NovaSeq data are not likely to be an artefact of the DADA2 analysis. Moreover, it indicates that extremely deep sequencing is required if one wants to have a comprehensive survey of biodiversity in a region.

The MiSeq is less capable of sequencing low-abundance eDNA. Our NovaSeq results indicate that in the locations in which we sampled, approximately 9,300 FishE ESVs are present (Fig. 5). However, the MiSeq was only able to obtain approximately 3,500 ESVs even at an unrealistically-high sequencing depth (Fig. 4). This suggests that the MiSeq could not identify approximately 60% of the diversity present in the environment. In order to determine taxonomic/biological breadth of these ESVs we performed taxonomic assignment on all the ESVs from both instruments—i.e., we combined the results of both MiSeq runs to give that platform the best possible chance of finding all the taxa present—and found that the MiSeq data contained 80 identifiable families. The NovaSeq also identified these same 80 families but was also able to identify an additional 32—a 40% increase. Those families unique to the NovaSeq analysis are listed in Table 2. Some of the taxa missing from the MiSeq data are of significant interest, including marine mammals (Delphinidae) and several fish. Other taxa include those that are clearly not marine organisms (e.g., cow and moose) but this is not surprising given the sampling sites' close proximity to a human-populated shore, and still demonstrates that organisms with presumably low-abundance eDNA are less likely to be detected by the MiSeq than the NovaSeq.

Kingdom	Phylum	Class	Order	Family	Common name
Metazoa	Annelida	Polychaeta	Capitellida	Capitellidae	Bristleworm
Metazoa	Annelida	Polychaeta	Capitellida	Maldanidae	Bristleworm
Metazoa	Annelida	Polychaeta	Eunicida	Dorvilleidae	Bristleworm
Metazoa	Annelida	Polychaeta	Phyllodocida	Glyceridae	Bloodworm
Metazoa	Annelida	Polychaeta	Terebellida	Terebellidae	Marine bristleworm
Metazoa	Arthropoda	Collembola	Entomobryomorpha	Entomobryidae	Springtail
Metazoa	Arthropoda	Collembola	Entomobryomorpha	Isotomidae	Springtail
Metazoa	Arthropoda	Hexanauplia	Cyclopoida	Cyclopidae	Copepod
Metazoa	Arthropoda	Hexanauplia	Harpacticoida	Ameiridae	Copepod
Metazoa	Arthropoda	Hexanauplia	Harpacticoida	Normanellidae	Copepod
Metazoa	Arthropoda	Insecta	Diptera	Pediciidae	Crane fly
Metazoa	Arthropoda	Insecta	Hymenoptera	Ichneumonidae	Wasp
Metazoa	Arthropoda	Insecta	Lepidoptera	Noctuidae	Moth
Metazoa	Arthropoda	Insecta	Odonata	Coenagrionidae	Damselfly
Metazoa	Arthropoda	Malacostraca	Euphausiacea	Euphausiidae	Krill
Metazoa	Arthropoda	Ostracoda	Podocopida	Trachyleberididae	Seed shrimp
Metazoa	Bryozoa	Gymnolaemata	Cheilostomatida	Hippothoidae	Moss
Metazoa	Bryozoa	Stenolaemata	Cyclostomatida	Crisiidae	Moss
Metazoa	Chordata	Actinopteri	Beloniformes	Adrianichthyidae	Ricefish
Metazoa	Chordata	Actinopteri	Clupeiformes	Clupeidae	Herring
Metazoa	Chordata	Actinopteri	Cypriniformes	Cyprinidae	Carp
Metazoa	Chordata	Actinopteri	Pleuronectiformes	Scophthalmidae	Turbots
Metazoa	Chordata	Mammalia	Cetacea	Delphinidae	Atlantic white-sided dolphin
Metazoa	Chordata	Mammalia	Artiodactyla	Bovidae	Cow
Metazoa	Chordata	Mammalia	Artiodactyla	Cervidae	Moose
Metazoa	Cnidaria	Hydrozoa	Anthoathecata	Tubulariidae	Tubular hydroid
Metazoa	Cnidaria	Hydrozoa	Leptothecata	Campanulariidae	Jellyfish
Metazoa	Mollusca	Bivalvia	Pectinoida	Pectinidae	Scallop
Metazoa	Mollusca	Gastropoda	Littorinimorpha	Rissoidae	Minute sea snails
Metazoa	Mollusca	Gastropoda	Trochida	Margaritidae	Sea snails
Metazoa	Nemertea	Palaeonemertea		Cephalothricidae	Ribbon worm
Metazoa	Porifera	Demospongiae	Dendroceratida	Halisarcidae	Skeletonless sponge

Table 2. List of families detected with the NovaSeq but not with the MiSeq. Many are biologically significant taxa, including dolphins and several fish.

Interestingly, when these taxa are plotted on a circular dendrogram we do not observe any obvious phylogenetic pattern to the distribution of missing taxa on the MiSeq (Fig. 6). Rather, it appears that the NovaSeq was generally able to detect more families within each order than the MiSeq.

Although we have no quantitative measurements of the abundance of taxa present in the locations we sampled, we note that many of the taxa missing from the MiSeq analysis are likely to have a very low abundance of eDNA (e.g., marine mammals, terrestrial organisms) compared to taxa where whole organisms or gametes may be present in the water samples (e.g., zooplankton). We can approximate this by looking at read abundances (Fig. 7). If we assume that read abundances roughly correlate with the original biomass present¹⁷ then it does indeed seem that the MiSeq is less capable of sequencing this low abundance eDNA than the NovaSeq, even at very high sequencing depths.

Whether or not this phenomenon will have a significant impact on a particular experiment depends strongly on the purpose of the experiment. Those that are attempting to detect or catalogue rare or endangered species may be strongly impacted, since these are precisely the organisms that the MiSeq is likely to miss. On the other hand, general biodiversity assays or comparative studies in community composition will be less impacted by missing these low-abundance taxa. To illustrate the point, we generated NMDS plots for genera identified within the 9 samples (3 sites) that were deep-sequenced on the MiSeq and NovaSeq. Even though the NovaSeq detected more genera overall, qualitatively both instruments pick up a gradient change from coastal to deeper waters along the primary axis (Fig. 8).

Discussion

Most environmental metabarcoding studies are not sequencing deep enough. Our results suggest that using seawater as the source of environmental DNA at a typical sequencing depth of 60,000 reads per sample, only half of the diversity detectable by the MiSeq will be captured (Fig. 2). To reach the MiSeq's detection limits for analysis of seawater one would have to aim for 0.8–1 million reads per sample per marker—more than ten times the typical depth of sequencing currently performed in most metabarcoding studies. We further note

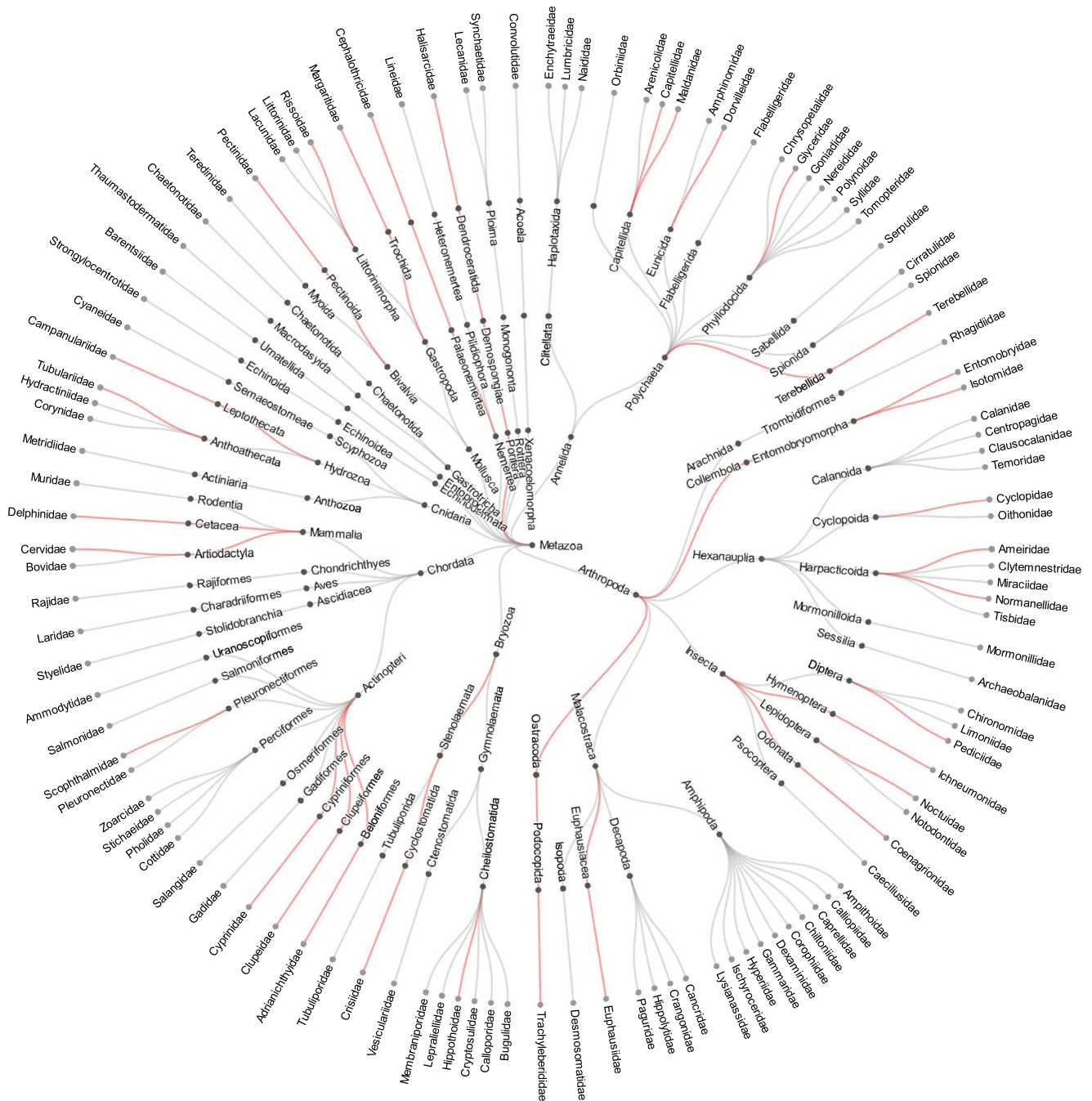


Figure 6. Radial dendrogram of taxa identified in our experiments. The leaves of the tree represent family-level taxa. Grey edges on the tree indicate taxa that both the NovaSeq and MiSeq platforms detected. Edges in red were only detected on the NovaSeq; there were no taxa unique to the MiSeq.

that the samples from our study were obtained from the North Atlantic Ocean where biodiversity is presumably far less than that which might be found in tropical regions. For this reason, even deeper sequencing may be required in regions or substrates that have very high biodiversity.

Even when matched for depth, the NovaSeq can detect greater diversity than the MiSeq. The most remarkable finding in this study is that the NovaSeq can detect many taxa that the MiSeq cannot—even when the depth of sequencing is matched. This is true on a PCR -by-PCR basis (Fig. 1), and even greater depth of sequencing on the MiSeq cannot overcome this obstacle (Fig. 4). The outcome is that there may be a great deal of missing biodiversity in MiSeq analyses (Fig. 6).

Whether or not this has a significant impact on a study will depend on the nature of that study. The inability to detect low-abundance taxa is unlikely to have a large impact on comparative community composition analyses. On the other hand, studies that have an interest in low-abundance taxa (e.g., those that are rare or endangered)

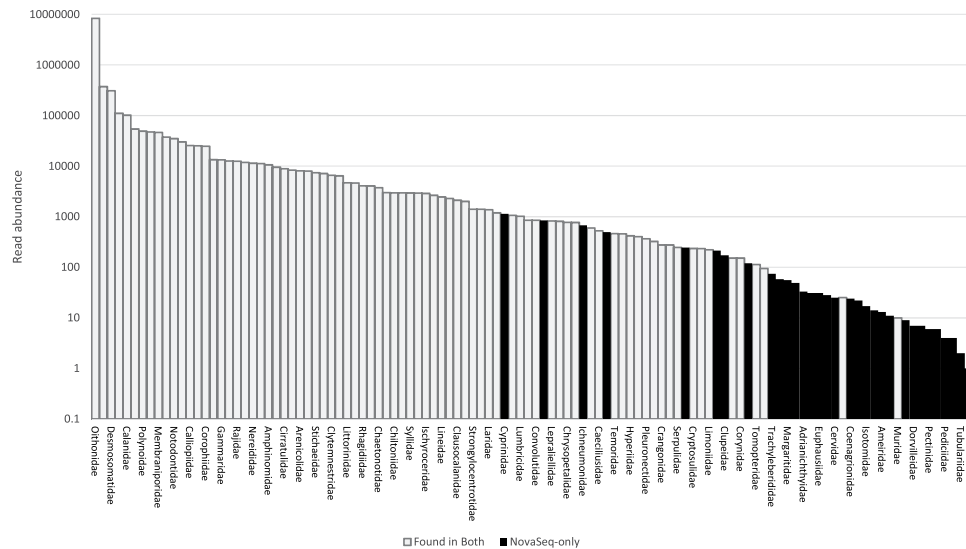


Figure 7. Families detected on the NovaSeq ranked by read count (note the y-axis has a logarithmic scale). The white bars indicate taxa that were detected by both the NovaSeq and the MiSeq, while black bars indicate taxa detected solely by the NovaSeq. We note that most of the taxa missed by the MiSeq have low read abundance.

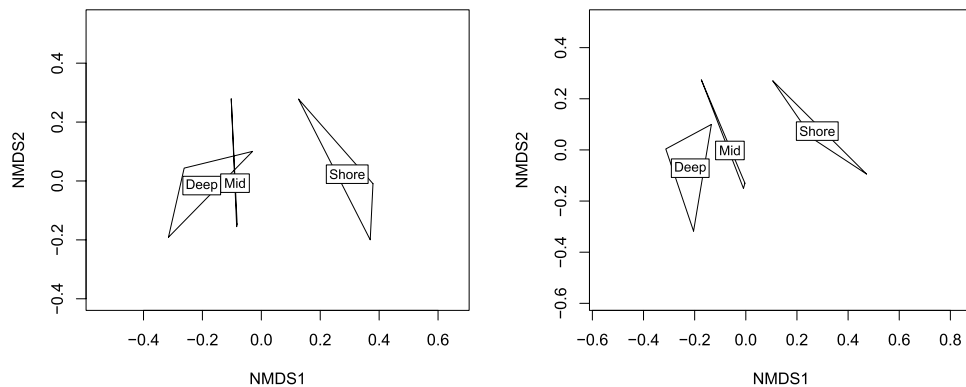


Figure 8. NMDS plots of genera from the MiSeq data (left) and the NovaSeq data (right) show that there is little qualitative difference between the two.

could be very significantly impacted. We also note that many studies comparing eDNA-based approaches to traditional morphological methods frequently show eDNA-based methods missing taxa that were observed visually. We wonder if some of these cases could be explained through the lower sensitivity of the MiSeq to low-abundance eDNA.

Possible causes for the difference between platforms. The NovaSeq has many differences from the MiSeq, including: a 2-colour chemistry rather than 4-colour used by the MiSeq; greatly improved hardware (presumably including better image capture abilities); different signal processing software; and a flow cell that has pre-defined binding spots for target DNA instead of the random lawn used by previous Illumina instruments. We suspect but cannot be certain that this last factor—the flow cell—is a significant cause for the NovaSeq's superior performance. In the MiSeq and most previous Illumina instruments, DNA binds to the flow cell in a random fashion. Therefore, to distinguish one spot on the flow cell where DNA has bound from another, the spots are observed by the instrument for the first 25 rounds of sequencing and at that point clusters are determined¹⁸. This works well when performing shotgun sequencing (the primary use of Illumina's instruments) because the spots can be clearly distinguished from each other thanks to the high level of sequence diversity. However, when performing amplicon-based sequencing the variability from one spot to the next—especially within the first 25 bases which likely covers primer regions—is minimal and this can cause two distinct spots to be merged together. To prevent this from happening, Illumina recommends spiking in PhiX genome¹⁸, but unless the proportion of PhiX is very high it's nearly impossible to prevent similar sequences from sitting near each other on the flow cell. Conversely, the spots that DNA anneal to on the NovaSeq flow cell are pre-defined and known by the instrument's base calling software, so inferring their location is not necessary and this largely prevents the “over-clustering” of low diversity reads.

Notably, the MiSeq runs that we performed for this project used the recommended levels of PhiX and the sequencing run statistics all matched Illumina recommendations. Nevertheless, we still suspect that over-clustering is at least partially responsible for the MiSeq missing out on diversity that the NovaSeq was able to detect.

Our results suggest that the NovaSeq 6000 may be a superior instrument for environmental metabarcoding studies especially in complex biodiversity-rich substrates where heterogenous abundance of taxonomic groups may confound detection. However, while the MiSeq is a relatively inexpensive instrument that could plausibly be obtained by many labs, the NovaSeq is expensive—roughly ten times the cost of the MiSeq. Moreover, while it produces hundreds of times more data per run, each run is also significantly more expensive: the smallest currently available NovaSeq flow cell and sequencing kit is approximately ten times more expensive than the popular MiSeq v3 600-cycle kit. For these reasons, the NovaSeq may be out of reach for many laboratories in the near term and we therefore suggest a few approaches that may aid obtaining more comprehensive biodiversity from the MiSeq.

Multiplex different markers together on the same flow cell. This is already quite a common practice, albeit frequently for money-saving purposes rather than to prevent over-clustering. In theory, multiplexing several markers (while still maintaining adequate sequencing depth per sample), will lead to greater sequence diversity on the flow cell and will reduce the probability of over-clustering.

Use dual-indices on the sequencing primers. It is common practice to add short (e.g., 8mer) oligonucleotide indices to the sequencing primers for the purposes of multiplexing, but this has the added benefit of contributing to additional base composition diversity on the flow cell. It should be noted that we did employ a dual-index strategy in the present paper (see the Methods), so this strategy alone does not seem sufficient to close the gap between the MiSeq and NovaSeq instruments.

Use large amounts of PhiX spike-in. PhiX genome fragments also serve to increase the complexity on the flow cell, but many labs try to minimize the amount of PhiX they use because it takes up precious sequencing capacity. Paradoxically though, increasing PhiX may in fact increase the number of quality reads generated. Illumina's own recommendations range from 5–50%¹⁹ although in practice most experiments end up at the lower end of this range.

Use phased amplicons. Another approach was suggested by²⁰, who designed overlapping 16 s amplicons that they described as “phased amplicon sequencing”. Despite covering the same region of interest, the different sequence composition at the 5' end of the read reduced over-clustering by the MiSeq—so much so that the number of reads passing quality filters increased by 9–47% in their experiments.

Conclusions

Biodiversity analysis through genomics has enabled widespread applications from human microbiome studies to environmental assessment and monitoring. With rapid advances in sequencing hardware and computational approaches for data analysis, it is important to determine the impact that sequencing technology and strategy have on the data generated, especially where it may influence biological interpretations and their socio-economic implications. Here, we tackled the issue of sequencing technology and depth on an analysis of biodiversity in seawater through eDNA metabarcoding. Our analysis provides direct evidence of the superior utility of the newly introduced NovaSeq platform for elucidating a more comprehensive biodiversity measurement as compared to the current workhorse, the MiSeq platform. Our results strongly suggest that comprehensive detection of biota from eDNA in a complex environment such as the ocean is possible and will aid supporting scientific/societal endeavours for enhanced biodiversity analysis for people and the planet.

Methods

Sample collection. Triplicate 250 mL water samples were taken from surface water simultaneously. Samples were taken from eight locations along two transects in Conception Bay, Newfoundland and Labrador, Canada, on October 13–14, 2017 (Fig. 9). These samples cover a range from near-shore to approximately 10 km offshore (with a sea bottom depth ranging from a few metres nearshore to approximately 200 metres in the middle of the bay).

Laboratory procedures. *DNA extraction.* Filtration and DNA extraction was done in PCR clean laminar flow hoods (AirClean Systems) thoroughly decontaminated with ELIMINase (Decon Labs) and 70% EtOH prior to each set of three sample replicates. Water samples were thawed at 4 °C and immediately filtered with 0.22 µm PVDF Sterivex filters (MilliporeSigma). The DNeasy PowerWater Kit (Qiagen) was used to extract DNA with the automated QIAcube platform (Qiagen), following the DNeasy PowerWater IRT protocol. For lysis, bead tubes were heated for five minutes at 65 °C and then vortexed for ten minutes. Negative controls were generated during filtration and extraction to screen for contamination and cross-contamination. Filtration and extraction were done in a pre-PCR room isolated from post-PCR rooms.

DNA library preparation. Two fragments were amplified by PCR from the 5' end of the standard COI barcode region: the 235 bp F230 fragment¹⁰, and the 232 bp Mini_SH-E fragment²¹. Illumina-tailed PCR primers (tails underlined) were used to amplify targets: The F230 forward primer (LCO1490; 5'-TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG GGT CAA CAA ATC ATA AAG ATA TTG G-3',²²), the Mini_SH-E

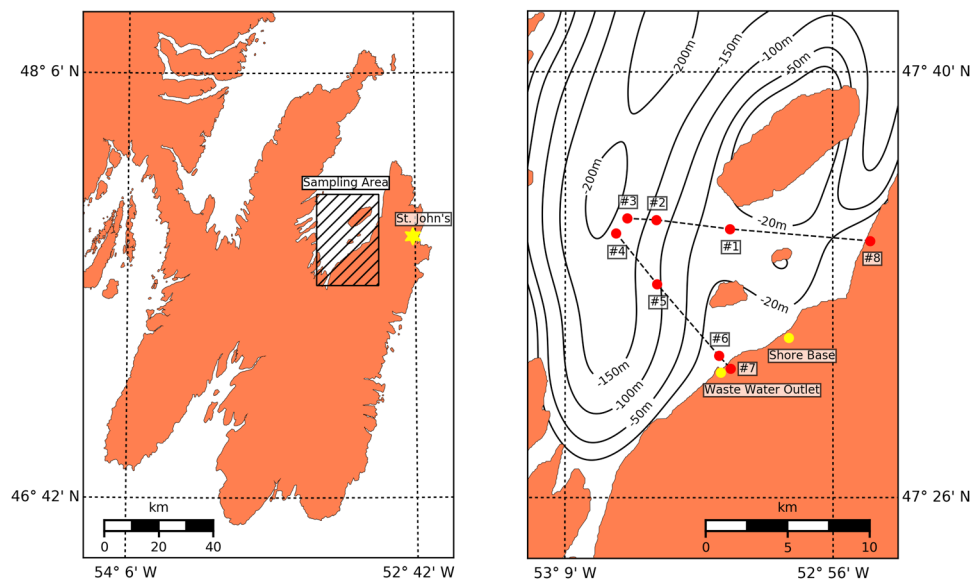


Figure 9. Location of the eight sampling sites from Conception Bay, Newfoundland and Labrador, Canada.

forward primer (Fish_miniE_F_t; 5'-TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG ACY AAI CAY AAA GAY ATI GGC AC-3'), and the reverse primer (230_R/Fish_miniE_R_t; 5'-GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GCT TAT RTT RTT TAT ICG IGG RAA IGC-3') was used for both F230 and Mini_SH-E fragments. Each amplification reaction contained 0.6 μ L DNA, 1.5 μ L 10X reaction buffer, 0.6 μ L MgCl₂ (50 mM), 0.3 μ L dNTPs mix (10 mM), 0.3 μ L of each Illumina-tailed primer (10 μ M), and 0.3 μ L Platinum Taq (0.5 U/ μ L; Invitrogen) in a total volume of 15 μ L. PCR conditions were initiated with a heated lid at 95 °C for 3 mins, followed by 35 cycles of 94 °C for 30 s, 46 °C for 40 s, and 72 °C for 1 min, and a final extension at 72 °C for 10 mins. Three PCR replicates were amplified from each sample with the ProFlex thermocycler (Thermo Fisher) and then pooled for a single PCR cleanup with the QIAquick 96 PCR purification kit (Qiagen; 60 μ L elution volume). Agarose (1.5% w/v) gel electrophoresis was used to verify amplification of samples, and for quality control of negative controls from PCR, extraction, filtration, and field collection. Each indexing reaction contained 2 μ L amplicon DNA, 2.5 μ L 10X reaction buffer, 1 μ L MgCl₂ (50 mM), 0.5 μ L dNTPs mix (10 mM), 1 μ L of F indexing primer (5 μ M), 1 μ L R indexing primer (5 μ M), and 0.5 μ L Platinum Taq (0.5 U/ μ L; Invitrogen) in a total volume of 25 μ L. Unique dual Nextera indexes were used to mitigate index misassignment (IDT; 8-bp index codes). PCR conditions were initiated with a heated lid at 95 °C for 3 mins, followed by 12 cycles of 95 °C for 30 s, 55 °C for 30 s, and 72 °C for 30 s, and a final extension at 72 °C for 5 mins. Indexing success was verified on the Bioanalyzer (Agilent) with the DNA 7500 kit. Samples were quantified with Quant-iT PicoGreen dsDNA assay with a Synergy HTX plate fluorometer (BioTek) and pooled to normalize DNA concentration. Libraries were cleaned with three successive AMPure XP cleanups: Left side selection with bead:DNA ratios of 1 \times , then 0.9 \times , and a right-side selection with 0.5 \times . Libraries were quantified with a Qubit fluorometer (Thermo Fisher) and the size distribution was checked with the DNA 7500 kit on the Agilent 2100 Bioanalyzer. Two libraries containing F230 or FishE amplicons from nine samples were sequenced on the Illumina MiSeq with two 600-cycle v3 kits. Two libraries containing F230 or FishE amplicons from 24 samples were pooled with other libraries and sequenced with two MiSeq. 600-cycle v3 kits. Field, filtration and extraction negatives were also sequenced in these MiSeq runs. Two libraries containing F230 and FishE amplicons were sequenced with a 300-cycle S4 kit on the NovaSeq 6000 following the NovaSeq XP workflow.

Bioinformatics. We employed two different workflows to analyze our data in order to reduce the possibility that our results were an artefact of the method used. In both workflows, base calling and demultiplexing were performed using Illumina's bcl2fastq software (version 2.20.0.422). Primers were then trimmed from the forward and reverse reads using cutadapt v1.16²³ with the default error tolerance and a minimum overlap equal to half the primer length. We discarded read pairs in which the primer was missing from either the forward or reverse read. We note that both amplicons studied in this paper are very short (~230 bases) and therefore there is ample overlap between the forward and reverse reads in both the MiSeq (300 cycle forward and reverse) and NovaSeq (150 cycle forward and reverse) data. We saw no significant difference in the rate of successful paired-end joining between the two instruments.

After this stage the two methodologies diverged and are described separately below.

DADA2 workflow: DADA2 v1.8.0¹⁵ was used to perform quality filtering and joining of paired reads (maxEE = 2, minQ = 2, truncQ = 2, maxN = 0), and denoising (using default parameters) to produce exact sequence variants (ESVs). This was performed independently on the MiSeq and NovaSeq data since their error patterns are presumed to be different and therefore they require different models to be trained. Singleton ESVs

were discarded. To rapidly evaluate the overlap in ESVs between the two instruments, MD5 hashes²⁴ were generated for each of the ESV sequences and then these sets of hashes were compared between the MiSeq and the NovaSeq.

OTU clustering workflow: Vsearch v2.8.2²⁵ was first used to join the paired ends of the reads (using default parameters), and perform quality filtering (using default parameters). The reads for the NovaSeq and MiSeq were then dereplicated, and these reads were combined into a single file so that OTU clustering (using the cluster_fast setting) could be performed on the entire set using an identity threshold of 97%. As with the ESVs, singleton OTUs were discarded.

Taxonomic assignment: NCBI's blastn tool v2.6.0²⁶ was used to compare ESV sequences against the nt database (downloaded August 2018), using an e-value cut-off of 0.001. We filtered the resulting hits with the requirement of having at least 90% identities across at least 90% of the query sequence. In cases where there was not a single unambiguous best hit, we used a majority consensus threshold of 80% to assign taxonomy²⁷.

Accumulation curves: Original read memberships were tracked through the various analytical steps: dereplication followed by OTU clustering, or ESV generation using DADA2. Subsamples were then generated using sampling proportional to the original read abundances with the "choices" function within the Python programming language's "random" module²⁸. These reads were then mapped to their respective ESVs/OTUs for comparison between the two DNA sequencing platforms.

NMDS plots: NMDS plots were generated using the default settings of the metaMDS function, part of the vegan library²⁹ in the statistical package R³⁰. Data were based on genera in both rarefied NovaSeq and whole MiSeq data that could be identified with 95% or better identity across 95% or more of the read.

Data Availability

All data have been deposited into NCBI's Sequence Read Archive under accession number PRJNA513845.

References

- Aylagas, E., Borja, Á. & Rodríguez-Ezpeleta, N. Environmental Status Assessment Using DNA Metabarcoding: Towards a Genetics Based Marine Biotic Index (gAMBI). *PLOS ONE* **9**, e90529 (2014).
- Baird, D. J. & Hajibabaei, M. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol. Ecol.* **21**, 2039–2044 (2012).
- Veldhoen, N. *et al.* Implementation of Novel Design Features for qPCR-Based eDNA Assessment. *PLOS ONE* **11**, e0164907 (2016).
- Aylagas, E., Borja, Á., Muxika, I. & Rodríguez-Ezpeleta, N. Adapting metabarcoding-based benthic biomonitoring into routine marine ecological status assessment networks. *Ecol. Indic.* **95**, 194–202 (2018).
- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* **21**, 2045–2050 (2012).
- Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proc. Biol. Sci.* **270**, 313–321 (2003).
- Ratnasingham, S. & Hebert, P. D. N. bold: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Mol. Ecol. Notes* **7**, 355–364 (2007).
- Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C. & Baird, D. J. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One* **6**, e17497 (2011).
- Ji, Y. *et al.* Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett.* **16**, 1245–1257 (2013).
- Gibson, J. F. *et al.* Large-Scale Biomonitoring of Remote and Threatened Ecosystems via High-Throughput Sequencing. *PLoS One* **10**, e0138432 (2015).
- Shaw, J. L. A. *et al.* Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. *Biol. Conserv.* **197**, 131–138 (2016).
- DiBattista, J. D. *et al.* Assessing the utility of eDNA as a tool to survey reef-fish communities in the Red Sea. *Coral Reefs* **36**, 1245–1252 (2017).
- Cahill, A. E. *et al.* A comparative analysis of metabarcoding and morphology-based identification of benthic communities across different regional seas. *Ecol. Evol.* **8**, 8908–8920 (2018).
- Hajibabaei, M., Baird, D. J., Fahner, N. A., Beiko, R. & Golding, G. B. A new way to contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. *Phil Trans R Soc B* **371**, 20150330 (2016).
- Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
- NovaSeq. 6000 System Quality Scores and RTA3 Software Available at: <https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/novaseq-hiseq-q30-app-note-770-2017-010.pdf> (2017).
- Evans, N. T. *et al.* Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Mol. Ecol. Resour.* **16**, 29–41 (2016).
- Optimizing Cluster Density on Illumina Sequencing Systems Available at: <https://www.illumina.com/content/dam/illumina-marketing/documents/products/other/miseq-overclustering-primer-770-2014-038.pdf> (2016).
- How much PhiX spike-in is recommended when sequencing low diversity libraries on Illumina platforms? Available at <https://support.illumina.com/bulletins/2017/02/how-much-phix-spike-in-is-recommended-when-sequencing-low-diversity.html>. (Accessed: 14th November 2018)
- Wu, L. *et al.* Phasing amplicon sequencing on Illumina Miseq for robust environmental microbial community analysis. *BMC Microbiol.* **15**, (2015).
- Shokralla, S., Hellberg, R. S., Handy, S. M., King, I. & Hajibabaei, M. A. DNA Mini-Barcoding System for Authentication of Processed Fish Products. *Sci. Rep.* **5**, 15894 (2015).
- Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* **3**, 294–299 (1994).
- Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
- Rivest, R. RFC1321: The MD5 Message-Digest Algorithm. Available at: <https://www.ietf.org/rfc/rfc1321.txt> (1992).
- Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- Schloss, P. D. & Westcott, S. L. Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *Appl. Environ. Microbiol.* **77**, 3219–3226 (2011).
- Van Rossum, G. Python Programming Language. In *USENIX Annual Technical Conference* (2007).
- Oksanen, J. *et al.* vegan: Community Ecology Package Available at: <https://CRAN.R-project.org/package=vegan> (2018).

30. R Core Team. R: A Language and Environment for Statistical Computing. (2018). Available at: <https://www.R-project.org/>.
31. Abrams, J. F. *et al.* Shifting up a gear with iDNA: from mammal detection events to standardized surveys. *bioRxiv* 449165 <https://doi.org/10.1101/449165> (2018).
32. Ando, H. *et al.* Evaluation of plant contamination in metabarcoding diet analysis of a herbivore. *Sci. Rep.* **8**, 15563 (2018).
33. Burgess, T. I., McDougall, K. L., Scott, P. M., Hardy, G. E. S. & Garnas, J. Predictors of *Phytophthora* diversity and community composition in natural areas across diverse Australian ecoregions. *Ecography* **42**, 565–577 (2019).
34. Cahoon, A. B., Huffman, A. G., Krager, M. M. & Crowell, R. M. A meta-barcoding census of freshwater planktonic protists in Appalachia – Natural Tunnel State Park, Virginia, USA. *Metabarcoding Metagenomics* **2**, e26939 (2018).
35. Egeer, B. *et al.* Challenges for assessing vertebrate diversity in turbid Saharan water-bodies using environmental DNA. *Genome* **61**, 807–814 (2018).
36. Gran-Stadniczeňko, S. *et al.* Protist Diversity and Seasonal Dynamics in Skagerrak Plankton Communities as Revealed by Metabarcoding and Microscopy. *Journal of Eukaryotic Microbiology* Preview available at: <https://onlinelibrary.wiley.com/doi/abs/https://doi.org/10.1111/jeu.12700>.
37. Holman, L. E. *et al.* The detection of novel and resident marine non-indigenous species using environmental DNA metabarcoding of seawater and sediment. *bioRxiv* 440768 <https://doi.org/10.1101/440768> (2018).
38. Hugoni, M. *et al.* Spatiotemporal variations in microbial diversity across the three domains of life in a tropical thalassohaline lake (Dziani Dzaha, Mayotte Island). *Mol. Ecol.* **27**, 4775–4786 (2018).
39. Kerdraon, L., Balesdent, M.-H., Barret, M., Laval, V. & Suffert, F. Crop residues in wheat-oilseed rape rotation system: a pivotal, shifting platform for microbial meetings. *bioRxiv* 456178 <https://doi.org/10.1101/456178> (2018).
40. Nuske, S. J. *et al.* The endangered northern bettong, *Bettongia tropica*, performs a unique and potentially irreplaceable dispersal function for ectomycorrhizal truffle fungi. *Mol. Ecol.* **27**, 4960–4971 (2018).
41. Phan, H. C., Wade, S. A. & Blackall, L. L. Is marine sediment the source of microbes associated with accelerated low water corrosion? *Appl. Microbiol. Biotechnol.* **103**, 449–459 (2019).
42. Pochon, X., Wecker, P., Stat, M., Berteaux-Lecellier, V. & Lecellier, G. Towards an in-depth characterization of Symbiodiniaceae in tropical giant clams via metabarcoding of pooled multi-gene amplicons <https://doi.org/10.7287/peerj.preprints.27313v2> (2019).
43. Qian, X. *et al.* Shifts in community composition and co-occurrence patterns of phyllosphere fungi inhabiting *Mussaenda shikokiana* along an elevation gradient. *PeerJ* **6**, e5767 (2018).
44. Shahraki, A. H., Chaganti, S. R. & Heath, D. Assessing high-throughput environmental DNA extraction methods for meta-barcode characterization of aquatic microbial communities. *J. Water Health* **17**, 37–49 (2019).
45. Siegenthaler, A. *et al.* Metabarcoding of shrimp stomach content: Harnessing a natural sampler for fish biodiversity monitoring. *Mol. Ecol. Resour.* **19**, 206–220 (2019).
46. Too, C. C., Keller, A., Sickel, W., Lee, S. M. & Yule, C. M. Microbial Community Structure in a Malaysian Tropical Peat Swamp Forest: The Influence of Tree Species and Depth. *Front. Microbiol.* **9** (2018).
47. Vesterinen, E. J., Puisto, A. I. E., Blomberg, A. S. & Lilley, T. M. Table for five, please: Dietary partitioning in boreal bats. *Ecol. Evol.* **8**, 10914–10937 (2018).
48. Voulgari-Kokota, A., Grimmer, G., Steffan-Dewenter, I. & Keller, A. Bacterial community structure and succession in nests of two megachilid bee genera. *FEMS Microbiol. Ecol.* **95** (2019).
49. Wohlrab, S. *et al.* Metatranscriptome Profiling Indicates Size-Dependent Differentiation in Plastic and Conserved Community Traits and Functional Diversification in Dinoflagellate Communities. *Front. Mar. Sci.* **5** (2018).
50. Zinger, L. *et al.* Body size determines soil community assembly in a tropical forest. *Mol. Ecol.* **28**, 528–543 (2019).

Acknowledgements

The authors would like to thank Kirk Rees, Captain of the *Abrigo*, for assisting with sample collection. This work was partly funded through a Petroleum R&D Grant from InnovateNL (contract number 5405.2121.101), an award from the Atlantic Canada Opportunities Agency's Atlantic Innovation Fund (project number 781-37749-207993), and a grant from Petroleum Research Newfoundland and Labrador. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of Petroleum Research or its members.

Author Contributions

M.H. conceived the idea for this project and provided scientific oversight in experimental design, data analysis and interpretation and helped with writing the manuscript; N.A.F. designed and executed the sampling protocols; J.G.B. assisted with sampling and aided the bioinformatics analyses; A.M. led the laboratory work; G.S. performed the data analyses, figure generation, and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-42455-9>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019