JONAH VENTURES

Nanopore vs. Illumina metabarcoding

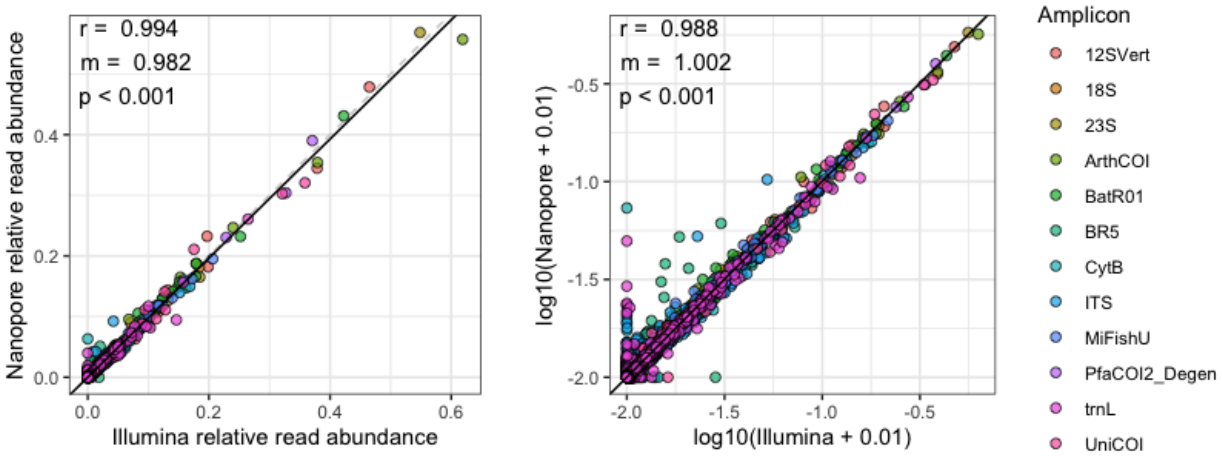
Nanopore GridION
(array of 5 MinION flow cells)

Illumina second-generation sequencing is the most common platform for DNA metabarcoding due to its historical high accuracy, sequencing depth, and throughput for short amplicons. In contrast, Oxford Nanopore Technologies (ONT) third-generation sequencing has been used less widely for short read sequencing because of concerns over high sequencing error rates. Given recent advances in the ONT platform that have reduced error rates and increased throughput, we compared the relative read abundance of exact sequence variants (ESVs) from 6 pairs of identical pools that were run on both an Illumina MiSeq and ONT MinION in order to determine whether the ONT platform can now be confidently used in metabarcoding studies.

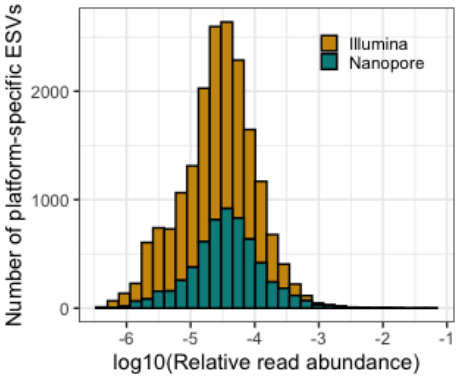
Methods: 1,614 water and animal fecal samples were extracted and PCR amplified for at least one of 12 primer sets, which targeted taxa across plants, fungi, vertebrates, and general eukaryotes. A total of 1,967 unique sample-amplicon combinations were prepared for sequencing across 6 pools by conducting a second round of PCR to append Illumina sequencing adapters and sample-specific index sequences. Sample library pools were sent for sequencing on an Illumina MiSeq at the Texas A&M Agrilife Genomics and Bioinformatics Sequencing Core facility using the v2 500-cycle kit (# MS-102-2003). The same sample pools used for Illumina sequencing were then prepared for ONT sequencing using the Ligation Sequencing Kit V14 (# SQK-LSK114). Pools were sequenced using an ONT 10.4.1 MinION flow cell at Jonah Ventures. Final basecalling was completed by running Nanopore's dorado software (v0.7.1) in super high accuracy mode using the dna_r10.4.1_e8.2_400bps_sup@v5.0.0 basecalling model.

Bioinformatics processing was the same for Illumina and ONT outputs. Cutadapt v3.4 was used to trim adapters and reorient reads in a 5' to 3' direction. The resulting reads were sorted into samples by demultiplexing with pheniqs v2.1.0, allowing no more than one mismatch in each of the paired 10 bp molecular indices. Cutadapt was then used again to extract the target amplicon by removing the primers, discarding any reads where one or both primers were not found or where the resulting amplicon sequences were outside of the expected length range. Exact sequence variants (ESVs) were then identified for each sample using the unoise3 denoising algorithm in vsearch. Only reads that were observed 4 or more times and with a maximum expected error rate > 1 bp were considered as ESVs. Chimeras were filtered using uchime3. Final read counts were determined for each sample by mapping unfiltered raw reads to the ESVs using usearch_global in vsearch, with a minimum identity of 95%. For statistical analysis, ESV relative abundance was averaged across samples for each sequencing run.

Results: Both sequencing platforms had similar numbers of processed reads per run, with an average of 5,915,093 for Nanopore and 5,840,094 for Illumina. Between the two platforms, a total of 38,176 unique ESV-by-run events were generated. Nanopore and Illumina relative read abundances for the ESVs were highly correlated across all runs and amplicons, with a Pearson r = 0.994 ($p < 0.001$) for untransformed data and r = 0.988 ($p < 0.001$) for $\log_{10}(x + 0.01)$-transformed data (*below*). In addition, there was no strong baseline shift in read abundance with either platform. For untransformed and transformed data, the slope of the reduced major axis line was near 1 (0.982 and 1.002, respectively) and the y-intercept near 0 (<0.001 and 0.005, respectively). Thus, as the ESV relative read abundance increased on one platform, the abundance increased on the other platform at nearly the same rate.



Of the 38,176 ESV-by-run events, 6,151 were unique to Nanopore (i.e., ESV reads were present in Nanopore and absent in Illumina) and comprised 1.2% of all Nanopore reads. Meanwhile, 13034 ESVs were unique to Illumina, comprising 1.7% of all Illumina reads. These unique ESVs were generally found in low abundance, with medians of 25 and 15 reads per unique ESV for Nanopore and Illumina, respectively (*right*). For 95% of ESVs, differences in relative read abundance between Nanopore and Illumina were within 3 per 1000 reads of each other. Yet, there was some variability across amplicons, with 95% of ESVs differing by 10 - 70 per 1000 reads for ArthCOI, CytB, BatR01, UniCOI, and PfaCOI2_Degen.



At the ESV-by-run level, maximum differences in relative read abundance ranged from 55 per 1000 reads greater on Illumina (38.0% vs. 32.6%) to 63 per 1000 reads greater on Nanopore (6.3% vs. 0%). Across all runs and amplicons, only 61 run-ESVs were over-represented in one platform by more than 10 per 1000 reads. For 48 of the 61 over-represented ESVs, there were no obvious complementary ESVs in the other platform to account for differences in relative read abundance (Table 1). These over-represented ESVs were more or less evenly distributed between platforms, with 22 over-represented in Nanopore (4 of which were absent in Illumina) and 26 over-represented in Illumina (all of which were present in Nanopore). The reason for these discrepancies in select ESVs is still unknown but may be attributable to general run-to-run stochasticity.

Table 1. Over-represented ESV in Illumina without complementary Nanopore ESVs to balance the difference.

ESV	Species	Mismatches	Nanopore	Illumina	Difference
007646	*Ursus americanus*	No similar ESVs	0.063	0.078	-0.015

The 13 remaining over-represented ESVs were found in the same run alongside other ESVs that shared at least 99% sequence similarity. When reads of these complementary ESVs were combined, the difference in relative read abundance between Nanopore and Illumina was reduced to just 2 - 7 per 1000 reads. When clustered together, the 13 ESVs were grouped into 5 clusters of highly similar ESVs: 3 of which were characterized by seemingly random substitutions of 1 - 2 nucleotides anywhere in the sequence (Table 2) and 2 of which by the variation of 3' homopolymers in Nanopore ESVs (Table 3). One example of random

Table 2. Similar ESVs over-represented in 1 of 2 platforms that, when combined, balance observed differences.

ESV	Species	Mismatches	Nanopore	Illumina	Difference
000297	*Odocoileus virginianus*	1	0.090	0.070	0.019
008037	*Odocoileus virginianus*	1	0.182	0.199	-0.017
			0.272	**0.269**	**0.003**

substitutions was found in the 12SVert amplicon for two ESVs that matched to *Odocoileus virginianus* (Table 2). In this case, the ESV that was more abundant in Nanopore had an adenine (A) at position 296 (of 395), while the ESV that was more abundant in Illumina had a thymine (T). When relative abundances were summed across the two ESVs, the difference in abundance of *O. virginianus* between platforms went from 0.019 and 0.017 for the ESVs to 0.003 for the taxon. Thus, collapsing ESVs by taxonomy or sequence similarity would likely reduce differences between platforms even further.

Table 3. Over-represented ESVs characterized by poly-A tails in Nanopore.

ESV	Species	3' tail	Nanopore	Illumina	Difference
000520	*Polygonum* sp.	GGAAGAAAAAAG	0.082	0.103	-0.021
010188	*Polygonum* sp.	GGAAGAAAAAG	0.014	0.000	0.014
194968	*Polygonum* sp.	GGAAGAAAG	0.012	0.000	0.012
			0.107	**0.103**	**0.005**

Both examples of homopolymer variation in Nanopore were found in the trnL amplicon. As an example, Illumina yielded only one ESV for *Polygonum* sp. and Nanopore detected that same ESV plus two others in lower abundance (Table 3). The only difference among the ESVs was the number of adenine (A) nucleotides at the 3' end (starting at position 120 of 124 - 127): 6, 5, and 3 - in the order of decreasing relative abundance. When the three Nanopore ESVs were summed, the relative read abundance of *Polygonum* sp. was comparable between Nanopore (0.107) and Illumina (0.103).

Conclusions: In comparing Nanopore and Illumina metabarcoding, we observed a strong correlation in relative read abundance between the two platforms. Both methods provided results that were largely indistinguishable. For ESVs with low read abundances, stochastic variation was more pronounced on the Illumina platform. Regardless, these unique and over-represented ESVs comprised a small percentage of the total reads in both platforms and likely would occur when comparing two runs of the same pool on the same platform. In all, both sequencing platforms yield functionally comparable outcomes for amplicon-based studies. Transitioning to Oxford Nanopore Technologies (ONT) sequencing will provide greater operational flexibility and faster results, as well as increased sequencing depth and the potential for longer amplicons, making it a robust advance in DNA metabarcoding.