Density based clustering and error correction of metabarcodes in Nanopore sequencing

Bilgenur Baloğlu[1], Zhewei Chen[2], Vasco Elbrecht[1,3], Thomas Braukmann[1], Shanna MacDonald[1], Dirk Steinke[1,4]

[1]Centre for Biodiversity Genomics, University of Guelph, Ontario, Canada
[2]California Institute of Technology, Pasadena, California, USA
[3]Centre for Biodiversity Monitoring, Zoological Research Museum Alexander Koenig, Bonn, Germany
[4]Integrative Biology, University of Guelph, Guelph, Ontario, Canada

Introduction

- Nanopore sequencing can enable field-based research, but few studies demonstrate its suitability for metabarcoding and analysis of environmental DNA
- We show metabarcodes in a bulk sample of 50 different aquatic invertebrate species can be identified with Nanopore Sequencing, and error corrected to accuracy comparable to MiSeq (up to 99.3% match against reference)
- Our python bioinformatics pipeline generates consensus reads from concatemers, performs OTU clustering with OPTICS, and enables exploration of error profiles and species composition
- Concatemer generation, error correction, and density based clustering enabled high fidelity identification and reconstruction of species barcodes de novo

Methods

- Mock sample preparation with Sanger sequencing
  - Tissue subsampling, DNA extraction, 658-bp COI amplification
  - 50 specimens with > 15% genetic distance selected for MiSeq and Nanopore experiment

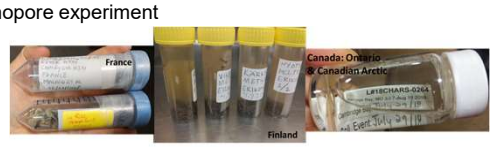

Figure 1: *Specimens in the mock sample*

- Metabarcoding with Illumina MiSeq 2x300

- Metabarcoding using Nanopore sequencing
  - Designed long UMIs (115 bp)
  - Rolling Circle Amplification (RCA): 2 experiments

Dataset	Protocol A	Protocol B
RCA duration (hrs)	5	6
Number of target sequences per RCA fragment	12	15
Enzymatic branching (min)	5	2
Mechanical fragmentation	4200 rpm, 2 min	None

Product	Version
Device	MinION MK1b
Flow Cells	MinION FLO-MIN/107.1 (R9.5)
Kits	Ligation 1D2
Data analysis	1D basecalling (not enough 1D2 data) ASHURE to process RCA reads and build consensus sequences (Baloglu et al, 2020, manuscript in review).

- Python-based bioinformatics pipeline (ASHURE)

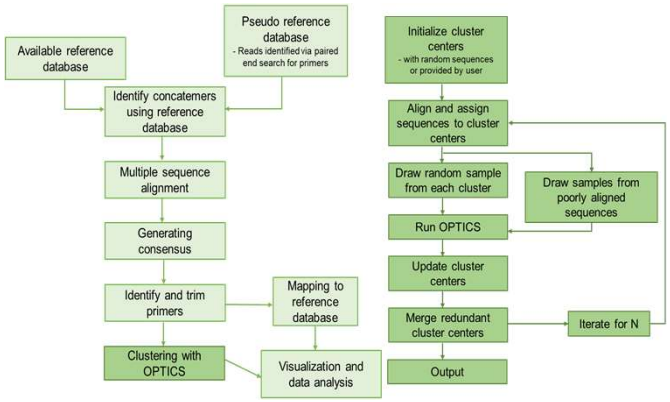

Figure 2: Overview of the ASHURE bioinformatics pipeline. Clustering step is colored in darker green and depicted in more detail on the right.

Results

- RCA was integral for increasing consensus accuracy, but was inefficient at generating long concatemers. Most RCA reads were still short (median fragment length of up to 1262bp)
- Median consensus accuracy of up to 99.3% was possible for long RCA reads (>45 barcodes)
- Consensus read error correlated better with de novo cluster center error, not so much with RCA length or UMI error
- ASHURE accuracy and sensitivity was comparable to MiSeq

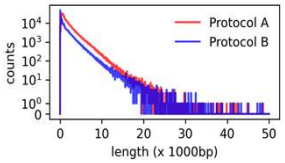

Figure 3: *Read length distribution*

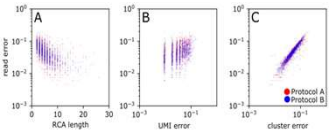

Figure 5: Comparison of consensus error versus (A) RCA length, (B) UMI error, and (C) cluster center error using the ASHURE pipeline for two RCA conditions.

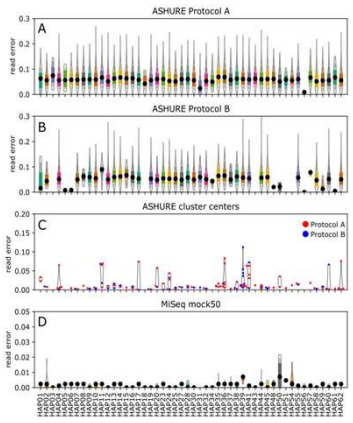

Figure 4: Nanopore sequencing read error per species for (A) Protocol A and (B) Protocol B obtained with ASHURE using all reads. (C) Nanopore sequencing read error obtained with OPTICS in ASHURE using cluster centers for each RCA condition. (D) MiSeq sequencing read error per species.

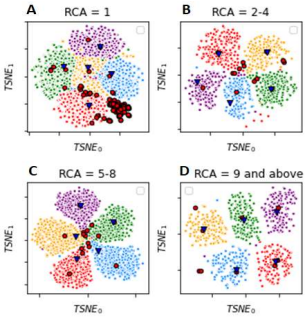

Figure 6: tSNE visualization of reference-free clustering using OPTICS for five randomly selected haplotypes with varying RCA fragment lengths. Blue triangles: True species obtained with Sanger. Red dots: De novo cluster centers.

Conclusions

- Nanopore sequencing can be used for metabarcoding with high accuracy (up to 99.3%)
- Density based clustering enables reference free OTU assignments for reads with mixed error profiles (https://github.com/BBaloglu/ASHURE)
- We recommend exploration of other isothermal amplification procedures for generating concatemeric reads
- This study was based on aquatic invertebrates, but the pipeline can be extended to many other taxa and ecological applications.

References

1- Baloğlu, B., Chen, Z., Elbrecht, V., Braukmann, T., MacDonald, S., Steinke, D. (2020). "A Workflow for Accurate Metabarcoding Using Nanopore MinION Sequencing." *bioRxiv*: 2020.05.21.108852.

CBG
CANADA FIRST RESEARCH EXCELLENCE FUND
APOGÉE CANADA FONDS D'EXCELLENCE EN RECHERCHE
UNIVERSITY OF GUELPH
FOOD FROM THOUGHT

bbaloglu@uoguelph.ca
@bilgeMolEcol