826

Assessing the impact of reference library completion on the temporal and spatial patterns of wetland communities identified through DNA metabarcoding

Michael TG Wright[1], Dr. Donald J Baird[2], Dr. Mehrdad Hajibabaei[1]
University of Guelph[1], Centre for Biodiversity Genomics[1], University of New Brunswick[2], Environment Canada @ Canadian Rivers Institute[2]

Abstract
Wetland macroinvertebrate samples were collected from Wood Buffalo National Park (Alberta, Canada) and were processed according to national biomonitoring standards, and for DNA metabarcoding. 73 taxa were morphologically identified, accounting for 1671 total observations. When limiting taxa identified through DNA metabarcoding to those same taxa and occurrences, DNA metabarcoding had an accuracy of 74.2%. This increased to 83.7% through the passive addition of sequences to GenBank over 18 months. This study highlights some potential targets in need of increased representation in genomic databases, including Harpacticoida, Collembola, and several mollusc orders. It also raises a question for consideration going forward with DNA metabarcoding in a biomonitoring context. If additions to genomic databases are able to incrementally change the outcome of studies over time, what can be done to minimize the risk of committing type 1 and type 2 errors in the assessment of ecosystem health (ie. saying a site is impacted when it is not, or failing to detect an impacted site.

Introduction

DNA metabarcoding allows for the rapid identification of taxa from bulk tissue samples, allowing us to bypass the need to subsample and sort complex environmental samples, such as benthic macroinvertebrate samples which are often used as indicators of ecosystem health[1,2]. The Biomonitoring 2.0 project in Wood Buffalo National Park provides an excellent opportunity to investigate important questions relating to the use of DNA metabarcoding such as:

How does DNA metabarcoding compare to taxonomic data obtained through typical morphological biomonitoring protocols? Accuracy? Patterns in community composition?

How do passive additions to genomic databases influence DNA metabarcoding data over time?

Methods

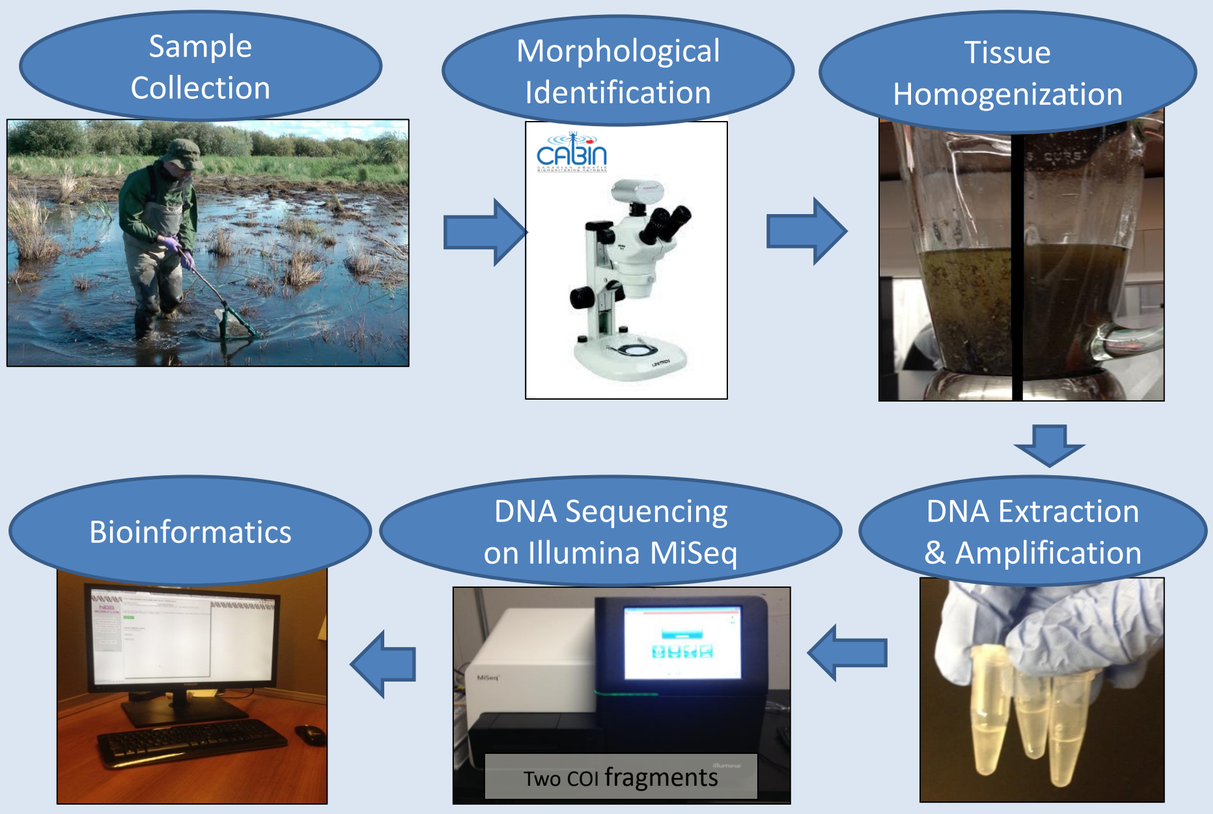

Fig 1. Schematic of processing workflow. Protocol outlined in Gibson *et al.* (2015)

78 wetland macroinvertebrate samples collected from the Peace-Athabasca delta (Alberta, Canada) in June and August of 2012 and 2013, following Gibson *et al.* (2015).

Taxonomy was assigned by comparing sequences against GenBank databases downloaded in September 2014 and April 2016.

Taxa identified through DNA metabarcoding were then compared against those identified morphologically to establish how well DNA metabarcoding represented the taxa known to be in the samples.

Results

73 invertebrate taxa were identified morphologically, accounting for 1671 total occurrences between all samples. When restricting DNA metabarcoding identifications to taxa identified morphologically, 1239 occurrences were observed using the September 2014 database (74.2% accuracy). This increased to 1381 occurrences when compared to the April 2016 database (82.7% accuracy). The largest discrepancies in DNA metabarcoding observations were found in Harpacticoida (0/46 morphological observations observed through DNA metabarcoding), Collembola (14/43), Physidae (16/66), and Planorbidae (53/73) (Fig. 2).

Variation in community composition was explained primarily by river delta of origin, and season of sampling (Fig. 3). Patterns of community composition did not significantly change between databases (Proctrustes rotation $p>0.05$).

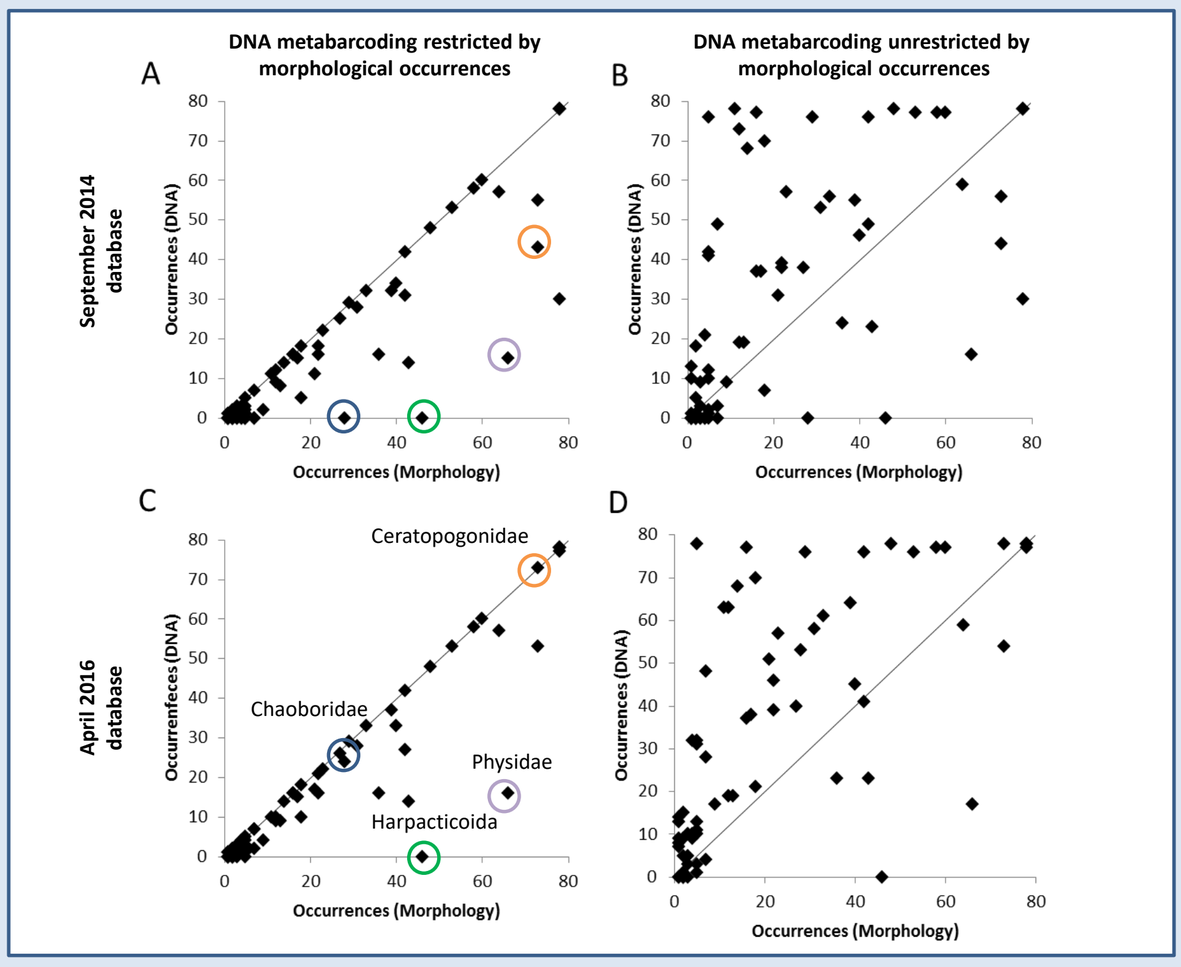

Fig 2. Scatterplot comparing morphological IDs to DNA metabarcoding IDs obtained from databases downloaded in September 2014 (A+B) and April 2016 (C+D). DNA metabarcoding IDs are either restricted to those where specimen was confirmed to be present (restricted, A+C) or unrestricted by presence in morphology (B+D)

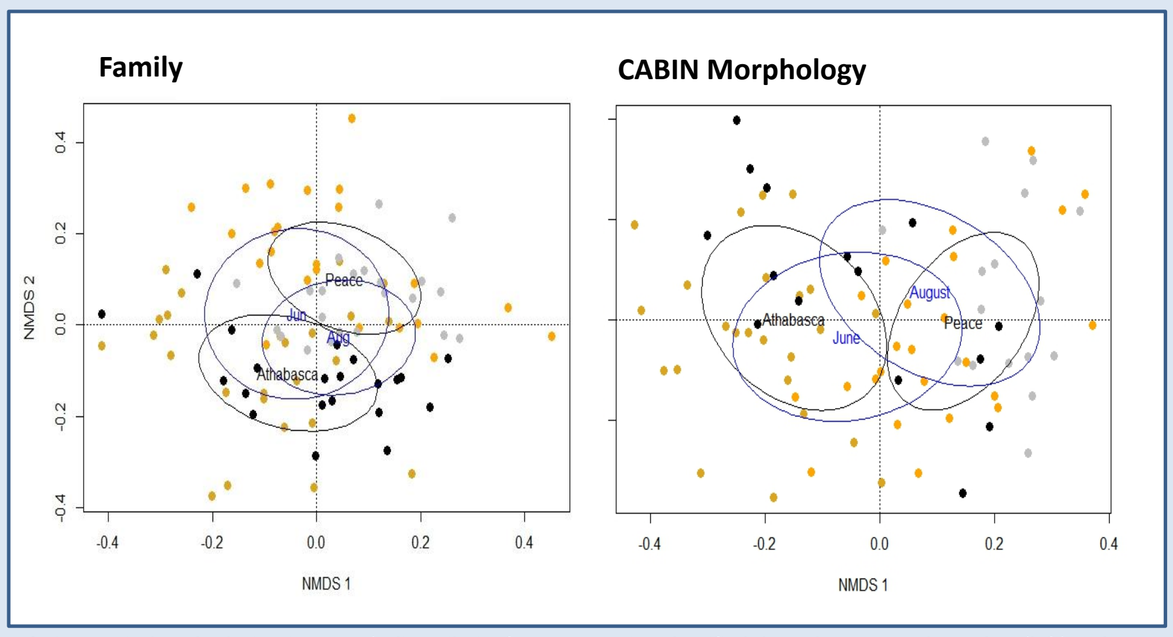

Fig 3. NMDS plots displaying Sorensen's dissimilarity for communities identified through DNA metabarcoding (left) and morphology (right). Ellipses represent standard deviation for river delta of origin, and sampling time.

Discussion and Future Directions

This study highlights some potential taxa which may be lacking adequate coverage in GenBank including: Molluscs, Copepods, and several non-insect arthropods. However, the increase in overall accuracy of 10.5% in 18 months is promising, given that this was through passive additions to the database (ie. no new sequences were uploaded from this project which would be expected to match closely). A next step would be to follow up with another update to the reference database, as well as delving deeper into into individual species coverage within GenBank.

Despite only 82.7% accuracy when comparing DNA metabarcoding directly to morphological observations, additional taxa not identified morphologically, and additional occurrences of observed taxa were identified through the bulk analysis of DNA metabarcoding, potentially indicating inadequate taxonomic representation when using the standard subsampling approach for morphology.

Additionally, despite differences in morphological and DNA metabarcoding identifications, community patterns are the same.

Consider the implications of an ever-increasing database:

- Can incremental changes to taxonomic inventories as a result of increasing database coverage cause us to make inaccurate ecosystem assessments (ie. Type 1 or 2 error)?
- If yes, how do you go forward with that knowledge?
- Not practical to repeatedly reanalyse data to incorporate changes to databases, especially when decision-makers have to decide whether or not to put resources into rehabilitation/protection, etc.

1. Barbour, M. T., J. Gerritsen, B. Snyder, and J. Stribling. 1999. Rapid bioassessment protocols for use in streams and wadeable rivers. USEPA, Washington.
2. Gibson, J. F., S. Shokralla, C. Curry, D. J. Baird, W. A. Monk, I. King, and M. Hajibabaei. 2015. Large-Scale Biomonitoring of Remote and Threatened Ecosystems via High-Throughput Sequencing. PLoS ONE **10**:e0138432.
This project was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute through the Biomonitoring 2.0 project (OGI-050) (to M.H.).

BIOMONITORING 2.0 · Parks Canada · Parcs Canada · Environment Canada · Environnement Canada · OntarioGenomicsInstitute · GenomeCanada · Centre for Biodiversity Genomics · University of Guelph