Supplementary Material

# All scripts from this section are available here: <https://github.com/Allaby-lab/PIA-accessories>

# Benchmarking methods detail

## Filtering by negative controls

All sequences from all negative controls in the sequencing run were combined into a single FASTA. The sequencing run contained 142 sediment samples including the seven used in this study. This combined FASTA was analysed with MEGAN and PIA alongside the seven samples, resulting in an “-ex” (extracted nodes) text file for the former and summary basic file for the latter.

A custom Perl script was used to filter all sample output files by the relevant control:

# perl Filter\_Summary\_Basic\_or\_MEGAN\_ex\_by\_control.pl [control] [sample]

The default threshold is 0.02, so taxa with a control:sample ratio of 0.02 or above were discarded from that sample. This excludes taxa that appear in negative controls unless they have a much greater presence in the sample. All samples were filtered individually and concatenated into a single file for visualising.

## Visualising with Krona

Another custom Perl script was used to convert the combined filtered output file to a format readable by Krona:

# perl Convert\_Summary\_Basic\_or\_MEGAN\_ex\_for\_Krona.pl [input] [output]

We produced the Krona chart using the default settings with two exceptions: -k to keep assignations to “cellular organisms”, and -t 1 to allow correct use of the .forKrona.txt files.

# ktImportTaxonomy [MEGAN forKrona] [PIA forKrona] -o [sample name].krona.html -k -t 1

We generated an SVG file using “Create snapshot” in the open HTML. For clarity, we adjusted the colours using the “Color Shift” function in Inkscape (version 0.92) specifying “Shift (°)” as 150 and “Saturation” as 100.

# Accuracy testing methods detail

The main text describes how the two sets of 250 test sequences were obtained from GenBank and analysed using BLAST, MEGAN and PIA. This section details the custom scripts used to re-format and interpret the input FASTAs and the results so that the sequences could be tracked to their final assignments. The Embryophyta FASTA is used in example code but the Mammalia FASTA followed the same process.

First came two steps to re-format the original FASTA.

Convert the default GenBank multi-line FASTA to single-line (no line breaks in sequences) using fasta\_formatter from the FAST-X toolkit 0.0.13 (Gordon & Hannon, 2010).

# fasta\_formatter -i 250\_sequences\_Embryophyta.fasta -o 250\_sequences\_Embryophyta.reformatted.fasta

Reduce the FASTA headers to just the identifier field. This version of PIA could not handle long query sequence names containing spaces.

perl Reduce\_FASTA\_headers\_to\_identifiers.pl 250\_sequences\_Embryophyta.reformatted.fasta

## Control analysis (original database)

We began with the more detailed BLAST that generates input for PIA:

# blastn -db [database] -num\_threads 10 -query -250\_sequences\_Embryophyta.reformatted.reduced.fasta -out 250\_sequences\_Embryophyta.reformatted.reduced.control.txt -max\_target\_seqs 500 -outfmt “6 std staxids”

The MEGAN analysis was performed manually. The BLAST and FASTA files were read into MEGAN5 using a GI mapping file to define taxonomy. This produced an RMA file. All nodes were exported twice in CSV format as “readName\_to\_taxonId” and “readName\_to\_taxonName”.

The PIA analysis used this command:

perl PIA.pl -f 250\_sequences\_Embryophyta.reformatted.reduced.fasta -b 250\_sequences\_Embryophyta.reformatted.reduced.control.txt -t 2

Note that PIA requires setup before first use. See the README for details (<https://github.com/Allaby-lab/PIA>).

We renamed the output directory to end in “\_control” for clarity.

## Exclusion analysis (reduced database)

The BLAST exclusion option, -negative\_gilist, requires the GIs (GenInfo Identifier) of every sequence to exclude. This would eventually be the GI of every sequence belonging to all 250 taxa. To start, we extracted the GIs of the 250 sequences:

perl Extract\_GIs\_from\_FASTA\_headers.pl 250\_sequences\_Embryophyta.fasta

Then we looked up the taxonomic IDs of those sequences using their GIs. This script uses the NCBI eUtils.

bash gi2taxid.sh 250\_sequences\_Embryophyta.GIs.txt

There is probably a way to use eUtils to download a list of GIs for all sequences belonging to a taxon, but we did this manually through the NCBI website. Then we concatenated the lists together:

 cat \*.txt > 250\_sequences\_Embryophyta.all\_GIs\_for\_taxa.txt

We then ran BLAST again excluding these GIs from the database:

# blastn -db [database] -num\_threads 10 -query -250\_sequences\_Embryophyta.reformatted.reduced.fasta -out 250\_sequences\_Embryophyta.reformatted.reduced.exclusions.txt -max\_target\_seqs 500 -outfmt “6 std staxids” -negative\_gilist 250\_sequences\_Embryophyta.all\_GIs\_for\_taxa.txt

The MEGAN analysis was performed manually as above.

The PIA analysis used this command:

perl PIA.pl -f 250\_sequences\_Embryophyta.reformatted.reduced.fasta -b 250\_sequences\_Embryophyta.reformatted.reduced.exclusions.txt -t 2

Note that, as above, we renamed the output directory to end in “\_exclusions” for clarity.

## Presenting results

One of the reference files used by PIA is an index matching taxonomic IDs to scientific names. This script uses that file to add names to the list of taxonomic IDs generated for the original FASTA:

perl id2name.pl [path to names.dmp.dbm] 250\_sequences\_Embryophyta.IDs.txt

We then merged this list of IDs and names with the identifier fields from the FASTA:

perl Extract\_read\_taxon\_FASTA\_and\_IDsnamed.pl 250\_sequences\_Embryophyta.reformatted.reduced.fasta 250\_sequences\_Embryophyta.IDs.named.txt

This script generates similar output from one of the two output files of PIA, the intersects file. We ran it on both the control and exclusion PIAs:

perl Extract\_read\_taxon\_intersects.pl 250\_sequences\_Embryophyta.reformatted.reduced.fasta.header\_out\_control/250\_sequences\_Embryophyta.reformatted.reduced.fasta.header\_out.intersects.txt

perl Extract\_read\_taxon\_intersects.pl 250\_sequences\_Embryophyta.reformatted.reduced.fasta.header\_out\_exclusions/250\_sequences\_Embryophyta.reformatted.reduced.fasta.header\_out.intersects.txt

Later versions of PIA will be able to output a file stating which read was assigned to which taxon without the need for a separate script.

Finally, we pasted the three read\_taxa files and the two pairs of MEGAN output files into a spreadsheet, sorted by read identifier, and matched read assignations together. Table S2 is the result.

# References

1. Entrez Programming Utilities Help. *National Center for Biotechnology Information (US)* (2010). Available at: https://www.ncbi.nlm.nih.gov/books/NBK25501/. (Accessed: 16th October 2019)

2. Gordon, A. & Hannon, G. J. Fastx-toolkit. (2010).