MIMP: MinIon Metabarcoding Pipeline

Geoffrey M. Williams © 2024

USDA Forest Service International Programs

Pipeline overview

Overview is that the approach is to use a taxonomic pre-assignment to divide up the work for mafft alignments (or pairwise) based distance matrices for clustering. The clusters are then re aligned to generate consensus sequences. Each step can be parameterized.

The pipeline for metabarcoding proceeds along the following process (note color codes).

Complete
In development
To be written

Guppy or Dorado, etc. (base-calling)
1 → *NanoFilt* (quality filtering)
2 → *cutadapt* (primer recognition and reorientation)\
3 → *MiniMap2* & *Samtools* (pre-alignment to database)
4 → *phyloseq/BioStrings/R* (filter by taxon and subset sequences
for within-taxon alignment and clustering)
5 → *mafft* (for lower taxons) → *phylip dnadist* (produce a distance matrix)
OR
5 → *mothur pairwise.dist* (distance matrices for higher level unknowns)
6 → *mothur cluster.classic* (cluster within the taxon alignments)
7 → TBA METHOD TO CHOSE CUTOFF
8 → *mothur bin.seqs*
9 → *mafft* (produce alignments within the subtaxon clusters)
10 → database of representative consensus sequences
11 → run *minimap2* on sequenes again to assign to refseqs

Important parameters currently supported/implemented in MIMP:

1. *trim_and_sort.sh*
    1. Nanofilt
        1. Quality cutoff (q)
        2. Minimum and maximum length (m, M)
        3. Head crop (c)
    2. Cutadapt
        1. Adapter error (e)
        2. Adapter overlap (o)
    3. Minimap
        1. Minimum sequence length (m)
2. *aggregate_minimap_bytaxon.sh*
    1. singleton cutoff (s)
    2. taxon level (l)
        1. rarefaction level (r)
3. *align_and_cluster_subtaxon.sh* – currently not parameterized but the following steps could all be parameterized
    1. mafft
    2. phylip dnadist
    3. mothur pairwise.dist
    4. mothur cluster.classic
    5. mothur bin.seqs
    6. consensus sequence calling

Scripts

1 *MIMP_README_MMYYYY.txt*
Detailed parameters for pipeline

2 *primer_seqs.sh*
Provides information on primer sets to calibrate script

3 *trim_and_sort.sh* (output file primer_pair/all_filt_reorient.fasta)

- i. runs NanoPlot (summary and visualization of quality and length, distribution etc. after basecalling)
- ii. runs NanoFilt (filters by q score, sequence length, and trims some reads like adapters from the reads)
- iii. converts file to .fasta format and shortens names of sequences in fasta sample headers for downstream compatibility
- iv. makes a new folder for the results for the primer set being used
- v. runs cutadapt to trim forward and reverse primer seqs and reorients them and contatenates all seqs from each barcode

4 *quick_dirty_minimap.sh* (output files all_filt.samview.tsv and relabund_phred_q10.csv)

i. minimap2 (alignment search for pre-classification) to database
ii. samtools view (process minimap2 output)
iii. *process_minimap2.R* (process samtools output)

5 *get_minimap_output.sh* (outputs folder e.g., minimap_ITS54/*.barcode*.csv)

6 *make_phyloseq.R* (outputs a .RData file with phyloseq object containing OTU table from minimap2)

7 *aggregate_minimap_bytaxon.sh* (pulls the sequences by specified taxon level and rarefies them in preparation for alignments, distance matrices, clustering, and consensus sequences)

i. *cluster_by_taxon_p1_parallel.R* (tabulates seqids by taxon and outputs them to a file)
ii. *cluster_by_taxon_p2_parallel.R* (Uses output from p2 to aggregate sequences at the desired taxonomic and rarefaction levels)

8 *align_and_cluster_subtaxon.sh* - This script is a dynamic command that performs several tasks to ulimately cluster and generate consensus sequences:

i. generates mafft alignment-based OR mothur pairwise.dist-based distance matrices
   i.a.1 mafft
   i.a.2 phylip dna.dist
   OR
   i.b mothur pairwise.dist
ii. from distances in (i), "subtaxon" clustering with mothur cluster.classic [NOTE: need to implement directory specification]
iii. "bin seqs" with mothur bin.seqs at the chosen distance cutoff label
iv. subsequent alignments of sequences within the subtaxon clusters and generation of consensus sequences

Note on basecalling

After initial basecalling, fast5_pass files were converted to pod5, and aggregated into a single pod5 file which was then re-basecalled (at super slow and accurate) and re-demultiplexed (at score cutoff 80) with the following parameters.