Read Me – Guidance to Aiden Protocol Scripts

This pipeline was created and optimised for Linux (Ubuntu 20). Each script file is saved in the accompanying folder and must be amended to include paths/file names as required. For this pipeline the following applications are required: amplicon\_sorter (https://github.com/avierstr/amplicon\_sorter), nbci blast (https://blast.ncbi.nlm.nih.gov/doc/blast-help/downloadblastdata.html), and python3.

**First**, for ONT data, if not already basecalled this can be done using dorado (https://github.com/nanoporetech/dorado). If raw files (pod5) are in subdirectories within a folder, the script *pods.sh* can be run first.

Dorado: ~1 hr

#this is for barcode kit SQK-RBK114-96. This value can be replaced or omitted.

./path/to/dorado basecaller hac ./path/to/pod5 --kit-name SQK-RBK114-96 > calls.bam

#this will basecall the pod5 files and can then be separated into barcodes by:

./path/to/dorado demux --emit-fastq --output-dir ./path/to/ouput --no-classify calls.bam

#this will create one file for each barcode as fastq files in the output folder

**Second**, the sequences can be filtered by quality score, normally removing scores of < 12

Script file: *filter1.sh* ~10 mins

**Third**, the amplicon\_sorter.py file needs to be in the folder with the Q12 sequences from above before running the script. This will create consensus and unique sequences from your reads and name them by barcode, gene, and predicted species (e.g., barcode01\_0\_0 is gene 0 from species 0). The read number of each will be in parentheses at the end of the fasta file name.

Script file: *amplicons.sh* ~hrs

**Fourth**, these sequences can then be blasted by remote tool (i.e., for any record on the online database, this requires an internet connection) which can be run in batches to cause less strain on the computing system

c) script file: *blastfin.sh*

These files can then be filtered to only include results which have query cover > 70 and sequence similarity > 97%.

Script file: *filter2.sh*

Then we may add the scientific name

Script file: *tax.sh*

And only retain unique references

Script file: *unique.sh*

The resulting file is then ready to be imported as a tab-delimited text file in R or any other program for analysis.