Jan 27, 2025 Version 4

Primary Data Analysis - Basecalling, Demultiplexing, and Consensus Building for ONT Fungal Barcodes V.4

DOI
https://dx.doi.org/10.17504/protocols.io.dm6gpbm88lzp/v4

Stephen Douglas Russell[1,2], Zachery Geurin[3], Josh Walker[4]
[1]Mycota Lab; [2]Biodiverse; [3]Tech Correct LLC; [4]Self

Mycota Lab

Stephen D Douglas Russell
Mycota Lab, Biodiverse, MycoMap

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.dm6gpbm88lzp/v4

Protocol Citation: Stephen Douglas Russell, Zachery Geurin, Josh Walker 2025. Primary Data Analysis - Basecalling, Demultiplexing, and Consensus Building for ONT Fungal Barcodes. **protocols.io** https://dx.doi.org/10.17504/protocols.io.dm6gpbm88lzp/v4Version created by **Stephen D Douglas Russell**

Protocol status: Working
We use this protocol and it's working

Created: March 07, 2024

Last Modified: January 27, 2025

Protocol Integer ID: 96268

Keywords: nanopore, bioinformatics, Specimux, demultiplexing, NGSpeciesID, MycoMap, consensus building for ont fungal barcode, ont fungal barcode, fungal barcode, barcode, fasta file, primary data analysis, usable fastq, minion run, data, raw fast5, fastq, basecalling, data analysis, containing consensus sequence, data from the run, protocol, consensus sequence

Abstract

This protocol assumes that your MinION run has been completed and the data from the run has been saved. It should take you from raw FAST5/POD5 read data to usable FASTQ/FASTA files containing consensus sequences for each of your fungal barcodes.

Note

This protocol assumes you are using 10.4.1 flowcells with V14 chemistry.

Before start

You should have POD5 or FAST5 files that are the result of a nanopore run which utilized tags from our previous protocol.

Shorthand Protocol

1 A shorthand version of this protocol can be found in the text document below. It contains summarized code to quickly run through the required bioinformatics steps, to be used once you are familiar with the process. Commands are written for Linux BASH terminal.

Dorado Run Code.txt 2.8KB

Note

If preferred, use a text editor to find-and-replace "ONT." (including the period) with your specific experiment name. For example: "ONT." -> "Run001." (ending with a period in both entries)

The commands, starting with BASECALLING, can then be copied into the terminal for ease-of-use. It is recommended to perform the initial preparation steps carefully to avoid issues.

Compute Setup

2 This protocol assumes that you have already installed all of the dependencies required from the master ONT DNA Barcoding Fungal Amplicons w/ MinION & Flongle protocol.

Note

It is best to restart your PC before continuing with the process, especially after completing a run in MinKNOW.

3 Download the following "Programs" folder, and unzip on the Desktop. You can leave it on your Desktop unzipped to copy into each new run as they occur.

Programs.zip 3KB

4 Download the latest version of the Specimux demultiplexing script from the GitHub repository: **[LINK]**

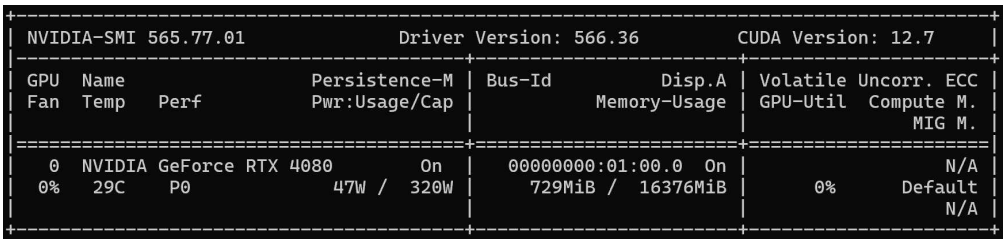
Right-click the link above, and choose "Save link as" to save the **specimux.py** file in the **unzipped Programs folder on the Desktop**.

Note

Specimux is in active development with updates planned throughout 2025. More information can be found on the GitHub **README**

5 **CUDA Troubleshooting:**
It's good to ensure your graphics card and CUDA driver are detected after initial install or when experiencing CUDA related errors from certain commands. Run the 'nvidia-smi' command, included in Linux **CUDA toolkit**. The terminal should print a table with the GPU and CUDA version, like so:

```
nvidia-smi
```



Note

For CUDA installations on Windows 11 (WSL2), download the **WSL2 CUDA toolkit** and refer to the documentation here: https://docs.nvidia.com/cuda/wsl-user-guide/index.html

Initial Post-Run, Pre-Analysis Preparation

6 **This protocol uses a generic "ONT" as the run name. Replace this with your specific experiment name, if preferred.**

Begin by opening a terminal (command line window).

Create a new folder on the desktop with the experiment name for processing the run data.
This is referred to as the "**working directory**".
Change directories to the new working directory.
Create a new folder in the working directory (named "pod5")

```
mkdir ~/Desktop/ONT
cd ~/Desktop/ONT
mkdir ./pod5
```

7 **[PREFERRED] If starting with POD5 format reads:**
It is best to select POD5 format output directly from MinKNOW during run setup.
Then copy all pod5 files from the MinKNOW data folder to the new pod5 folder.

7.1 **[PREFERRED] Option 1** : Drag and Drop in File Explorer. You can drag-and-drop or copy-and-paste the POD5/FAST5 folder/files of the run from the Minknow folder to your working directory.

The default directory in Linux is: "/var/lib/minknow/data/*run-name*/*cell-name*/*long_UID*/pod5/"

7.2 **Option 2**: Command Line. Be sure to change the various *names* to your specific folders (**HINT**: use **tab-completion** while typing within the linux command line to auto-populate names).

```
cp -r /var/lib/minknow/data/*run-name*/*cell-
name*/*long_UID*/pod5/ ./pod5
```

8 **[OPTIONAL] If starting with FAST5 format reads :**
If you are processing older data in the legacy FAST5 format, you must first convert the raw data to POD5.
Move the FAST5 files from the MinKNOW run folder into a fast5 folder within the new Desktop folder.
Then, convert the FAST5 files to POD5 files.

```
pod5 convert fast5 ./fast5/*.fast5 --output ./pod5/ --one-to-one
./fast5
```

9 Create an index file from your extraction template papers. This will allow you to link all of your reads with the individual specimens. A template for 10 plates (960 specimens) with ITS1F and ITS4 primers can be found here:

NANOPORE TEMPLATE SEVENTH ...

This .xlsx is formatted to utilize the Lab Code and iNaturalist # columns as the only inputs. It will combine these and all of the other columns into a single cell - concatenating them all into the final file name. For the Lab Code, we will typically put these into the iNaturalist "Voucher Number(s)" Observational Field, and then export them all at once into a .csv from iNat. This allows one to simply copy and paste many iNat numbers over at once, without ever needing to input any of the numbers manually, or use an XLOOKUP with the voucher number.

The spreadsheet can also be modified to accept Mushroom Observer (-MO) numbers or MyCoPortal occid numbers (-MP) instead of iNaturalist numbers (-iNat).

After editing, save as a tab-delimited text file in the Programs folder. You will need to remove most of the final columns from the template. The final output should be saved like this:

Index.txt

Basecalling with Dorado Simplex SUP mode

3h 15m

10 Run the Dorado basecalling command. The command below uses Dorado's simplex basecalling mode with the super-accuracy model. We do not do basecalling live with MinKnow because the latest algorithms are only available by running them standalone post-sequencing through the command line.

2h

```
dorado basecaller sup --no-trim ./pod5/ > ONT.raw.bam
```

For a Flongle cell with 1.15Gb of bases and 300 - 1.5M reads, this command can take 02:00:00 or more to run. Example output:

Expected result

```
user@pop-os:~/Desktop/ONT$ dorado basecaller sup --no-trim ./pod5 > ONT.raw.bam
[2024-11-24 18:57:38.193] [info] Running: "basecaller" "sup" "--no-trim" "./pod5"
[2024-11-24 18:57:38.205] [info]  - downloading dna_r10.4.1_e8.2_400bps_sup@v5.0.0
with httplib
[2024-11-24 18:57:42.586] [info] > Creating basecall pipeline
[2024-11-24 18:58:01.701] [info] cuda:0 using chunk size 9996, batch size 384
[2024-11-24 18:58:03.709] [info] cuda:0 using chunk size 4998, batch size 384
[2024-11-24 20:30:31.445] [info] > Simplex reads basecalled: 873596
[2024-11-24 20:30:31.445] [info] > Simplex reads filtered: 315
[2024-11-24 20:30:31.445] [info] > Basecalled @ Samples/s: 1.012127e+06
[2024-11-24 20:30:31.671] [info] > Finished
```

Preliminary QC Steps

3h 15m

11 Perform some initial length filtering on the results. The values in this step may need to be altered depending on the locus and primer combinations you are utilizing.

```
conda activate NGSpeciesID

# Full ITS - remove extra long and short reads (not ITS length)
(typically maintains Chanterelles)
samtools view -e 'length(seq)>400 && length(seq)<2000'  -O BAM -o
ONT.filtered.bam ONT.raw.bam

# Alternatively: sequencing ITS2 only - remove extra long and
short reads (not ITS length)
samtools view -e 'length(seq)>100 && length(seq)<700'  -O BAM -o
ONT.filtered.bam ONT.raw.bam
```

12 Convert the BAM files to FASTQ files for downstream compatibility.

```
samtools fastq ONT.filtered.bam > ONT.filtered.fastq

conda deactivate
```

Validate the QC of the Run

11h 20m

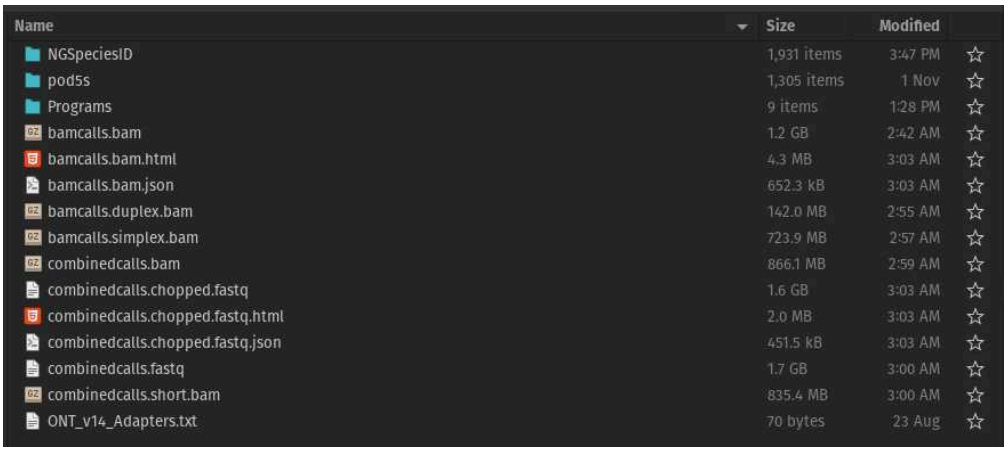
13 Compile quality control summary charts for the simplex and duplex reads.
Open the generated HTML files in any web browser to view QC reports
JSON files produced are not needed and can be removed, if preferred

```
conda activate sequali

sequali ONT.raw.bam
sequali ONT.filtered.fastq

conda deactivate
```

Expected result

Processing ONT.raw.bam: 100%|████████████████| 1.09G/1.09G [00:05<00:00, 233MiB/s]
Processing ONT.filtered.fastq: 100%|██| 1.53G/1.53G [00:02<00:00, 587MiB

You should now see a series of .html output files in your main directory. You can view the HTML files in a standard webbrowser. The JSON files are unused in this pipeline, but allow compatibility with the popular QC aggregator tool, **MultiQC**.



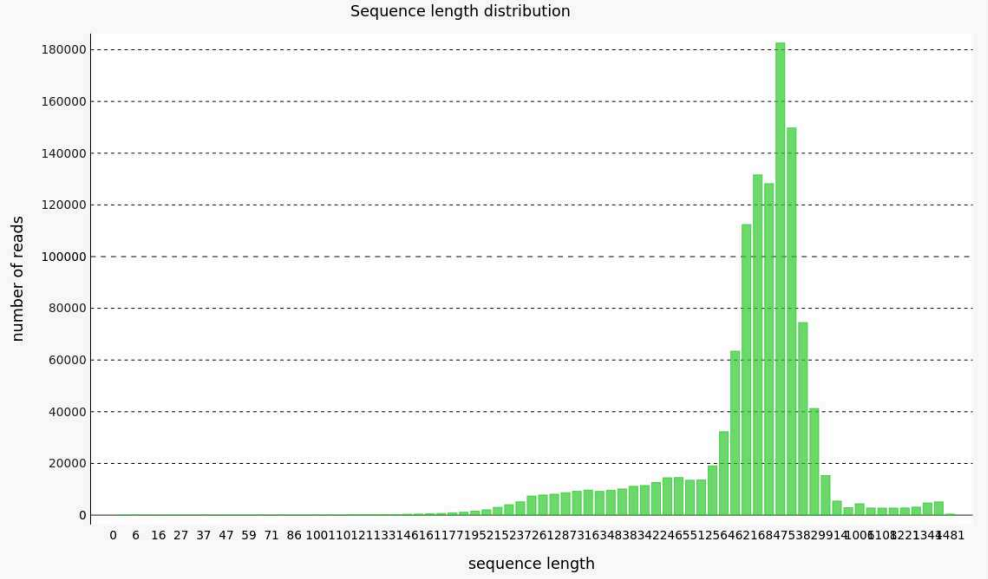
The Sequali QC step should generate several html and json files in your primary directory.

Review the images that are generated. Ensure the quality scores of your run are in an appropriate range. For a 10.4.1 Flongle with "Q20+" V14 chemistry, we typically get a peak in the 15-16 range, higher is better.

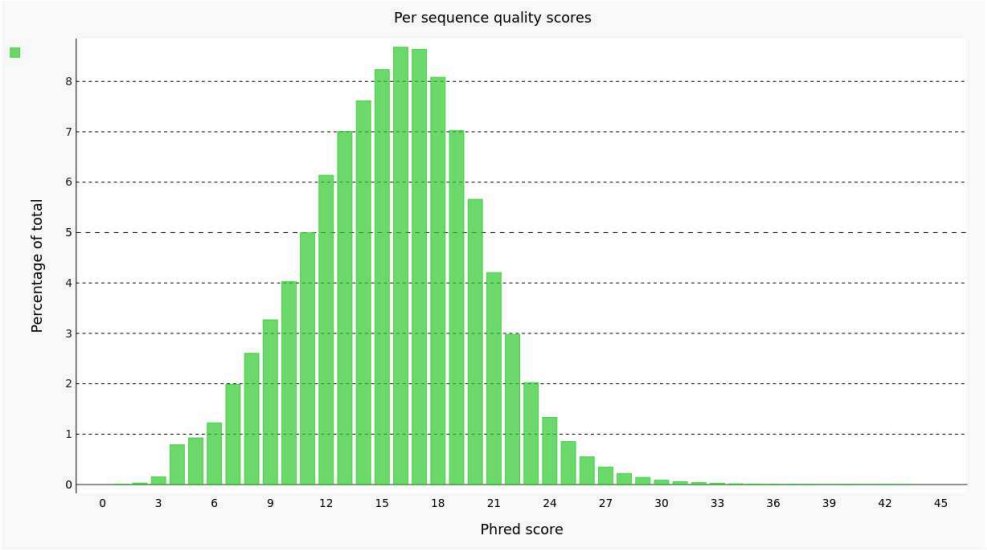
Summary

Mean length	669.91	
Length range (min-max)	1	1,499
Total reads	1,178,750	
Q20 reads	218,930	18.57%
Total bases	789,654,336	
Total GC bases	351,790,436	44.55%
Q20 bases	540,578,791	68.46%

The mean length should be appropriate for your target amplicon. Total read count should be near what MinKnow reported at the end of the run.



Should have a clear peak at your target amplicon length.



Q-scores should peak ~15 for a Flongle run.

Example of all outputs from this command:

combinedcalls.chopped.fastq.html 1.9MB

14 Perform some housekeeping to get your files and folder structures in place for the remainder of the protocol.

Note

IMPORTANT: Ensure your run's Index.txt file is available in the run directory and the unzipped Programs folder is on the Desktop (see Step 3 and 4 above). Note that the command line is case sensitive, including file and folder names.

```
mkdir NGSpeciesID
cp ~/Desktop/Programs/* ./NGSpeciesID/
cp ONT.filtered.fastq ./NGSpeciesID/
cp ./Index.txt ./NGSpeciesID/
cd ./NGSpeciesID
```

Demultiplex the Reads with Specimux

15 Demultiplex your samples using Specimux by providing the Index.txt sample key.

```
python specimux.py Index.txt ONT.filtered.fastq -F -e 3 -E 9 -d
```

At this step you should see ~20% - 40% of your reads properly demultiplexed into the correct buckets. If you are not in this range, you may need to double-check the primer and index combinations on your Index.txt sheet. If you are above this range, you have great PCR and libraries!

Expected result

```
user@pop-os:~/Desktop/ONT/NGSpeciesID$ python3 specimux.py Index.txt
ONT.filtered.fastq -F -e 3 -E 9 -d
2024-11-24 20:44:11,396 - INFO - Number of unique primer pairs: 1
2024-11-24 20:44:11,396 - INFO - Primer pair
GTGARTCATCGARTCTTTG/TCCTCCGCTTATTGATATGC: 768 specimens
2024-11-24 20:44:11,396 - INFO - Minimum edit distance is 6 for Forward Barcodes
2024-11-24 20:44:11,402 - INFO - Minimum edit distance is 5 for Reverse Barcodes
2024-11-24 20:44:11,409 - INFO - Minimum edit distance is 4 for Forward Barcodes +
Reverse Complement of Reverse Barcodes
2024-11-24 20:44:11,409 - INFO - Minimum edit distance is 12 for All Primers and Reverse
Complements
2024-11-24 20:44:11,417 - INFO - Minimum edit distance is 4 for Forward Barcodes +
Reverse Complement of Reverse Barcodes + All Primers
2024-11-24 20:44:11,417 - INFO - Using Edit Distance Thresholds 3 (barcode) and 9
(primer)
2024-11-24 20:44:11,420 - INFO - Will run 16 worker processes
read: 710883              matched: 191460       26.93%
2024-11-24 21:02:41,051 - INFO - Elapsed time: : 1109.63 seconds
2024-11-24 21:02:41,051 - INFO - Classification Statistics:
2024-11-24 21:02:41,051 - INFO - Matched                              : 191460 (26.93%)
2024-11-24 21:02:41,051 - INFO - No Barcode Matches                   : 157882 (22.21%)
2024-11-24 21:02:41,051 - INFO - No Reverse Barcode Matches (May be truncated) :
104943 (14.76%)
2024-11-24 21:02:41,051 - INFO - No Forward Barcode Matches (May be truncated) :
100190 (14.09%)
2024-11-24 21:02:41,051 - INFO - No Forward Barcode Matches           :  75755
(10.66%)
2024-11-24 21:02:41,051 - INFO - No Reverse Barcode Matches           :  47016
(6.61%)
2024-11-24 21:02:41,051 - INFO - Could Not Determine Orientation      :  26195
(3.68%)
2024-11-24 21:02:41,051 - INFO - No Reverse Primer Matches            :   5358 (0.75%)
2024-11-24 21:02:41,051 - INFO - Multiple Matches for Reverse Barcode :   1188
(0.17%)
2024-11-24 21:02:41,051 - INFO - No Primer Matches                    :    757 (0.11%)
2024-11-24 21:02:41,051 - INFO - No Forward Primer Matches            :    108 (0.02%)
2024-11-24 21:02:41,051 - INFO - Multiple Matches for Forward Barcode :     30
(0.00%)
2024-11-24 21:02:41,051 - INFO - Multiple Matches for Both Barcodes   :      1 (0.00%)
2024-11-24 21:02:41,051 - INFO - 76883 distinct barcode strings were unmatched
```

16 Remove several large files that are not necessary for most use cases and will just make your final analysis take longer.

```
rm sample_ambiguous.fastq
rm sample_unknown.fastq
rm ONT.filtered.fastq
```

Create the Final Consensus Sequences with NGSpeciesID

17 Utilize NGSpecies ID to generate your final consensus sequences from your demultiplexed samples. More info on NGSpeciesID can be found here: https://github.com/ksahlin/NGSpeciesID

```
conda activate NGSpeciesID

ls *.fastq | parallel NGSpeciesID --ont --consensus --t 1 --
abundance_ratio 0.2 --top_reads --sample_size 500 --
symmetric_map_align_thresholds --aligned_threshold 0.75 --
mapped_threshold 1.0 --medaka --fastq {} --outfolder {.}
```

Summarize the Data and Prep for MycoMap Upload

18 Create a summary file for your results by executing the summarize.py script within the NGSpeciesID folder. By default, consensus sequences with 2 or fewer Reads in Consensus (RiC) are filtered out of final summary folder. This threshold can be changed by adding "--min-ric N" to the command, where N is the minimum RiC **to be kept** (default = 3).

```
python summarize.py

# Using higher RiC threshold (filters RiC<=4; default is 3)
python summarize.py --min-ric 5
```

Expected result

```
Processing Medaka folder: sample_ONT08.90-B12-OMDL09960-
iNat192390863/medaka_cl_id_3...
Processing Medaka folder: sample_ONT08.90-B12-OMDL09960-
iNat192390863/medaka_cl_id_4...
Processing Medaka folder: sample_ONT08.90-B12-OMDL09960-
iNat192390863/medaka_cl_id_5...
Skipping /NGSpeciesID/sample_ONT08.95-G12-OMDL09965-
iNat61372973/medaka_cl_id_2/consensus.fasta due to low RIC (1 < 3)
Processing Medaka folder: sample_ONT08.96-H12-OMDL09966-
iNat61372916/medaka_cl_id_0...
Skipping sample_ONT08.96-H12-OMDL09966-
iNat61372916/medaka_cl_id_0/consensus.fasta due to low RIC (1 < 3)
Done!
```

19 Rename the Summary folder that is created to your experiment name (eg, "Run001_Summary") and compress it into zip file with the same name ("Run001_Summary.zip").

20 You are now ready to upload your summary folder into a MycoMap Project. Continue on with the secondary data analysis protocol.