Secondary Data Analysis - Creating a MycoMap Project from ONT Amplicon/Barcode Data V.2

Stephen Douglas Russell[1]

[1]The Hoosier Mushroom Society
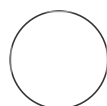
Stephen Douglas Russell

ABSTRACT

This protocol outlines many of the steps that can be taken after the primary data analysis to expedite the final analysis - turning your data into results.

Topics include creating a dashboard on MycoMap, automated BLAST searches and sequence error flagging, error resolution, linking sequences to iNaturalist observations, submission to GenBank, and many other topics.

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

Created: Sep 15, 2022

Last Modified: Mar 13, 2023

PROTOCOL integer ID:
70119

Validate Your Data

1 Before any further analytical work is done, it is always a good idea to validate the fidelity of your sequences to the iNat/MO records. As a part of the primary data analysis, you linked a spreadsheet of iNat/MO numbers to the unique primer tags for each specimen. Sometimes there are systematic errors in this process - primarily data entry errors or wrong associations with the tags on the spreadsheets. Typically, it is possible to figure out the origin of any systematic error and correct them.

In order to validate your data before upload to MycoMap, please BLAST at least two random sequences from each plate that was included in your nanopore run. NCBI BLAST will be the quickest way to do this. Do not utilize the first or last sequence from each plate (A1 or H12 well positions). If all of your data checks out, then you are fine to proceed with this protocol. If there are errors, you may need to BLAST more sequences to find the extent of the errors, fix them on your spreadsheet, and rerun the primary analytical steps so the sequences are associated with the proper observations.

I have yet to encounter a spreadsheet without some type of localized and/or structural error, so this is a vitally important step that will save you some time down the line.

Prepare your ONT Data

2 As a part of the "ONT Basecalling, Demultiplexing, and Analysis for Fungal Barcodes" protocol, you ran a python script called summarize.py. This created a folder called "__Summary__" in your working folder.

Change the name of this folder to the title you would like this run to be called. Ex - "First_Nanopore_Run" or "ONT.08.11.22." Compress this folder into a .zip file.

3 Depending on the number of samples in your final data, you will need to break it up into individual folders of not more than 250 sequences. This is because we will ultimately be BLASTing each sequence in the results. Running a BLAST on 1,000+ sequences with the NCBI API will take 36 - 48 hours. It also fills up the internal queue for the local BLAST for days. Consider breaking your initial folder from something like "ONT.08.11.22" to "ONT.08.11.22.1," "ONT.08.11.22.2", etc. Within each of these folders you will need to have a FASTQ folder with the appropriate files and you will need to copy and manually edit the summary.txt and summary.fasta files for each new folder.

Create a MycoMap Project

4 Login to MycoMap.com. If you are not a member:

Register for a username on MycoMap.com: https://mycomap.com/login

Then:

Create a project at the following link: https://mycomap.com/projects/create

Fill out the fields below:

Name: The name of your project. Something like "Mycoflora of Indiana - ONT001"
Description: You can enter text here that describes your project. This text will be shown in the header bar of the project, underneath the project title.
Project Type: Utilize the default of "Personal Working Project."
Parent Project: If you have other working projects, it may be beneficial to create parent and child projects. As an example, an "Mycoflora of Indiana - ONT001" project may flow into a parent "Mycoflora of Indiana" project. All of the records and validations included in the child project will flow into the parent project with this setting enabled. It can always be added at a later time.
MycoMap URL Path: This will generate a custom direct URL for your project. Ex- for a URL like http://www.mycomap.com/projects/**indianamushrooms** Enter: indianamushrooms into this field.
External Project URL: If you have an external website with information or data that describes your project, you can enter that URL here. The only thing this field does is to include the information you enter in the header of your project.
Taxon: If you only want a small group of mushrooms to be shown in this project by default, you can enter a specific taxon. Ex - "Amanita." It will include your selection in the defaults any time your project is opened. Typically, this field should be left blank. This is because if one of your records from iNat is labeled as "Amanitaceae," it would automatically be excluded from showing in your project by default.
**Location:** A with the taxon field, this allows you to enter a default location for your project. This field should typically be left blank, for the same reason as the taxon field should be left blank. Ex - if one of your iNat records has a hidden location, this record would be filtered out by default if a location is entered. You would have to do a special search in order to find it.
Owner: This option allows the project creator to select a MycoMap username that will be the owner of the project. If this field is left blank, the user who is logged in will be the owner of the project.
Enable List View: This is "checked" by default. The option allows non-members to see the Dashboard view. Non-members will not be able to manipulate any content, just view it.
Enable Files: This is "checked" yes by default. This allows the project to host file uploads/downloads. There are very limited use cases to have this turned off for your project.

Hit "Save" to create your project.
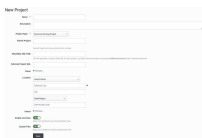
The new project creation screen on MycoMap.

Upload your ONT Summary File to the MycoMap Project

5 On the top-right of your MycoMap project, there are a series of green buttons seen below:

Primary project functionality within a MycoMap Project.

Click the green "Import" button.

6 Click on "Upload," then click on "Zip File of ONT FASTQ and Consensus Sequences." This will bring you to a screen with additional options for uploading ONT data into your project. Fill out the fields below:

Zip File: This is the zipped "Summary" file that was created as a part of the summarize.py script after running NGSpeciesID as a part of the "ONT Basecalling, Demultiplexing, and Analysis for Fungal Barcodes" protocol.

Import for Member: If you are creating a project for someone else, you can enter their MycoMap username here. Selecting a user here will create the MycoMap sequence files and BLAST searches under their name. This will allow that user to be the primary member who can manage the content.

Forward Primer: Enter the forward primer that was used for the amplicons with the ONT run. Ex - "ITS1F" Please contact the site administrator if your primers are not listed. Multiple primer combinations on a single run are not currently supported. They should be uploaded to projects as individual ZIP files.
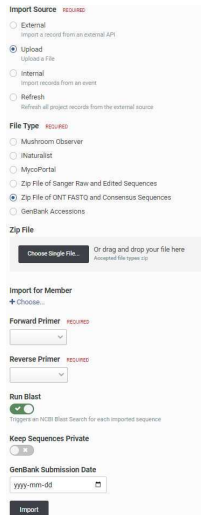
Reverse Primer: Enter the forward primer that was used for the amplicons with the ONT run. Ex - "ITS4" Please contact the site administrator if your primers are not listed. Multiple primer combinations on a single run are not currently supported. They should be uploaded to projects as individual ZIP files.

Run BLAST: This is "checked" by default. This setting will run a BLAST search and generate a BLAST results page for each sequence that was included as a part of the ZIP file.

Keep Sequences Private: This will make it so your sequences and raw data are not publicly viewable. There are limited use cases where this would be necessary.

GenBank Submission Date: If you want to flag your sequences with a target date for GenBank upload, it can be entered here.

Hit "Import" to import your data.

Note: There should is a check at the initial upload to ensure the ZIP file has a summary.fasta and summary.txt file. If either are missing, there is a popup message that does not allow you to proceed.


The import ONT ZIP file screen on MycoMap.

Backend Processing After Upload

7   Once this import is triggered, a large series of tasks are created that can take a while to complete. The length of time it takes to complete the tasks is determined by the current speed of the iNaturalist and NCBI API's, as well as the tasks from any other users on MycoMap that were put in before you. You should be able to see a progress task bar at the top of your MycoMap project, so you can track the progress of your upload. Contact steve@mycomap.com if your import does not appear to be regularly moving along.
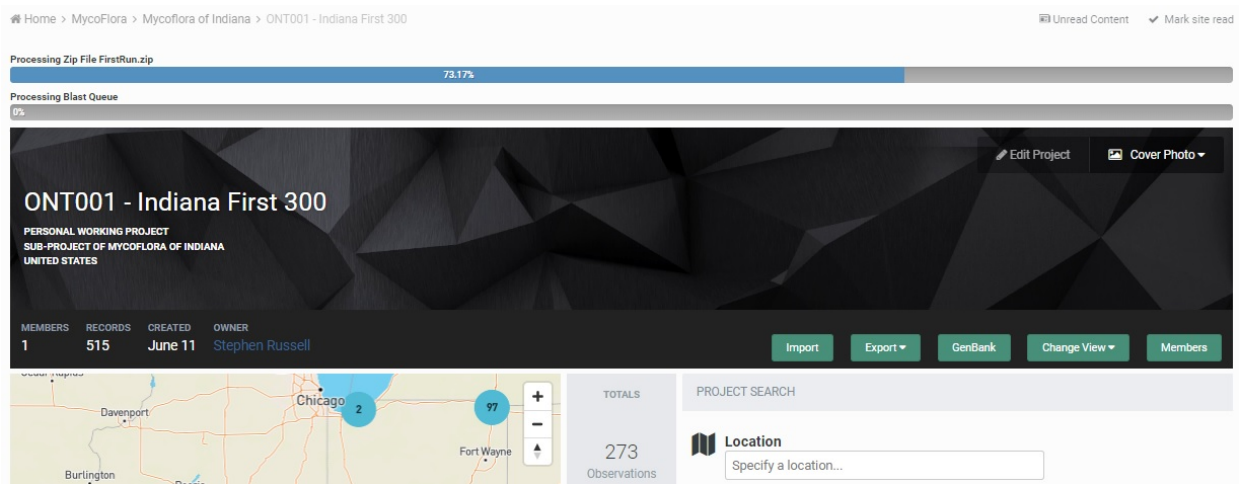
The blue status bar tracks progress of your upload processing. Depending on the size of your run, this could take several hours. Results will not be available until all BLAST searches have been completed. For this reason, it may be best practice to submit small batches of 100 sequences or less to the project at one time.

Lets take a quick overview of the things that occur once an upload is initiated. You will be able to track the progress of upload processing with a blue status bar at the top of your project.

Starting with the individual consensus sequences:

1. A MycoMap sequence record will be created for each sequence in your file. This is a database record within MycoMap that associates the consensus sequence with the raw data (FASTQ file), the RiC (Reads in Consensus), the iNaturalist or Mushroom Observer observation, and a MycoMap species name. An example MycoMap sequence record can be found here: https://mycomap.com/genetics/sequences/ont_sequences/off2021588-inat98620869-1-r25334 The sequence files are created in this publicly available "ONT Sequences" category (or they are hidden here if private was selected): https://mycomap.com/genetics/sequences/ont_sequences

The MycoMap sequence record page.

2. If the sequence has an RiC of 8 reads or less, then the sequence is flagged as being of particularly Low Quality and placed in a special "ONT Trashcan" category. This is an arbitrary cutoff that was selected to automatically remove sequences with this criteria from the manual review process. As a consensus with 10 reads or less is often associated with the wrong specimen (tag-switching), sequences in this category are not included in the local MycoBLAST results. Sequences can be moved into and out of the ONT Trashcan by utilizing the corresponding button at the top of each sequence record. Sequences with and RiC of 8-20 are included in the regular results, but flagged as Low Quality. The flag must be manually removed before the sequence can be uploaded to GenBank.

Button to move a sequence record into the trash and out of the MycoBLAST database.

Problem flags for ONT sequence records.

A line showing an individual iNaturalist observation on the MycoMap Dashboard. The "Sequence 'S'" can be seen in the middle in orange. Clicking this icon on each line with bring up the MycoMap sequence record that is associated with an observation.

3. There is sometimes more than one sequence associated with an individual observation as a result of this pipeline. The most I have seen is twelve. These are typically a result of (planned) bioinformatic processing or sequencing error, but can also be a result of lab contamination. In this case, assuming the RiC is above 20 for each sequence, a MycoMap sequence record is created for each sequence, and they are all associated with the observation on the dashboard. The number of these errant sequences can be reduced by playing with the allowable tag errors to cluster during primary processing. This can be altered to broaden or loosed with the MiniBar section of code. Changing e to zero will allow for zero errors in the primer sequence to form an association.

> **Command**
>
> **The default edit distance allowed between indexes (-e) is set to 4 base pairs and the edit distance allowed between primer sequences (-E) is set to 11 base pairs.**
>
> ./minibar.py -F Index.txt basecall.fastq -e 4 -E 11

4. If a consensus sequence is associated with more than one observation in the output (likely due

to sequencing errors within the tags [this is common]), then the resulting sequences are not included within the project and not associated with any individual observation. I have never found these to be useful, but they can be found in the original ZIP file.

5. If the top BLAST result to the sequence has a query coverage of less than 80%, then the sequence record is flagged as "Low Quality." There are a couple possibilities here. It is possible that there is still an A-tail or nanopore adapter at one or both ends of the sequence. These would need to be removed before GenBank upload. A sequence with the "low quality" flag will not be able to be uploaded to GenBank unless the flag is removed. This flag can be removed on the individual sequence record page or on the individual project line by clicking on the "S."

6. The title of the ZIP file that was uploaded is stored in the "Run Name" field associated with each sequence record.

7. If a sequence has more than two A's, C's, G's, or T's at either end, these are trimmed. This is because the pipeline will sometimes report sequences with a long tail remaining (Typically A or C). This is a lab artifact that should be removed.

8. For each consensus sequence record within the project, MycoBLAST search results are generated. This includes a standard NCBI (GenBank) BLAST as well as a local BLAST that houses all of the MycoMap-endemic data.

9. On the MycoBLAST results page, any record originating from an ONT/nanopore sequence has the ONT logo on the individual line, indicating that the result was from a nanopore upload. Typically, the remainder of the results would be from Sanger sequencing.

10. Using the results from the NCBI and local BLAST, a species name is assigned to the MycoMap sequence record. Details on changing this are found later in this protocol.

11. Using the NCBI BLAST results, if the first result has a negative orientation, a reverse-complement is applied to your sequence. The original sequence is also stored with the sequence record. It is common for the analytical pipeline to output sequences in the wrong orientation.

12. A revised summary.txt file is included within the Files section of the dashboard. The following columns should have been added for each sequence in the file:

(For each of these records, the metadata from the top result is reported. If the top result in the MycoBLAST is the same observation, then the metadata from the second result is reported)

NCBI BLAST Name
NCBI Identity
NCBI Query Coverage
NCBI Subject Coverage
Orientatation (Whether the subject coverage is + or -. [Plus or Minus]

MycoBLAST Name
MycoBLAST Identity
MycoBLAST Query Coverage
MycoBLAST Subject Coverage
MycoBLAST Orientatation (Whether the subject coverage is + or -. [Plus or Minus]
iNat/MO Number
Username

Problem Flags Summary for ONT Uploads.

8

PROBLEM FLAGS

- This sequence is a contaminant
- This sequence is the wrong specimen
- Low quality sequence
- Wrong Orientation

Problem flags that are utilized for ONT sequences.

This sequence is a contaminant - There are no automatic triggers for this flag. This flag can be manually utilized whenever the sequence results do not match the target sequence in an unexpected way. There are many ways a contaminant sequence can be expected. Ex - Hypomyces on a bolete. This is especially true when it is visible in the images of the associated observation. ONT sequences can report information on the target sequence, as well as any contaminants in the sample at the same time. As this flag removed the ability to add an individual sequence into GenBank, and you may want to upload expected contaminants into GenBank, only use this flag for unexpected, non-target contaminants.

This sequence is the wrong specimen - There are no automatic triggers for this flag. This flag should be used when the resulting sequence appears to be of good quality, but the result does not match up with the associated iNaturalist observation. There are many possibilities on how this could occur that would need to then be investigated. The wrong dried specimen could be in the assigned bag or there could have been a drying mixup. There could have been a mixup in the lab during the extraction and/or amplification process. Finally there could have been an error in generating the spreadsheet that associates individual specimens with their associated observations. Oftentimes, these errors can be elucidated with some investigation. Once the reason for the mixup is discovered, the sequence record should be associated with the right observation and this flag should be removed.

Low quality sequence - this flag is automatically triggered when the consensus sequence has an RiC of 20 reads or less, or the consensus has a top NCBI BLAST result query coverage of less than 80%. You may choose to use this flag manually when there is a basis to suspect a sequence is diverging from other reference data due to sequencing/bioinformatic errors, rather than biology. While it is rare for consensus seqs with a large RiC (200+) to have high error rates, it is occasional with RiC under 100, and uncommon in the 100-200 range. Manually setting this flag does two things. 1.) Removes the NCBI icon from the observation on the project dashboard. This prevents the sequence from being uploaded to GenBank. 2.) Places an "L" next to the sequence title on the local MycoBLAST.

Wrong orientation - this flag is automatically triggered when the top NCBI result has a different orientation than the consensus sequence.

Analytical Process

9   The top of the MycoMap project will have a status bar that tracks the progress of the sequence upload and subsequent BLAST searches. It would be prudent to wait until all of the tasks have completed, as the final summary spreadsheet is not able to be updated until all of the BLAST results are in. This may take quite a while, depending on how many sequences are a part of the run.


If there are individual sequences you are interested in, it is possible to search and view them on the dashboard (by species name or iNat number) while the other BLAST results finish. If the BLAST of your sequence is still pending, you can copy the sequence and BLAST through the NCBI interface.

10   The first thing I will typically do is to review the Summary.txt file that was created as a part of the process. An overall quick scan comparing the iNat species name with the top BLAST hits will allow systematic errors to be quickly identified and remedied. (Keep in mind that the summary file will only be fully updated once all of the BLAST searches for all of the sequences in the results have been completed. This is often why it is beneficial to break the total nanopore run down into subsets of 100 or less for upload. It may take a whole day for the NCBI API to process 1000 BLAST searches.)
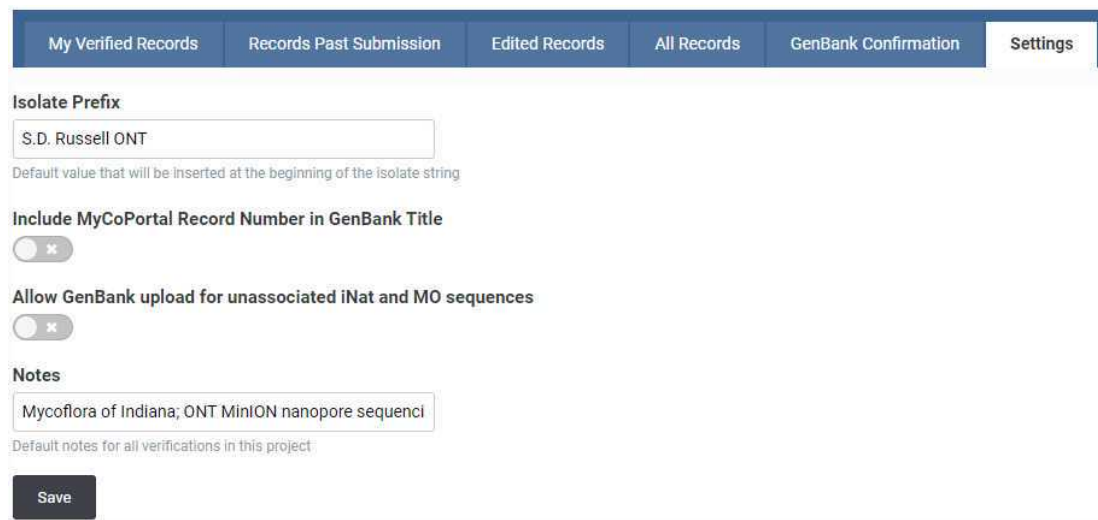
The summary file also gives me the opportunity to quickly share primary results for individual people or observations.

11 At the top of the MycoMap dashboard, there are a series of green buttons that perform a number of different tasks. Click on the "GenBank" button.
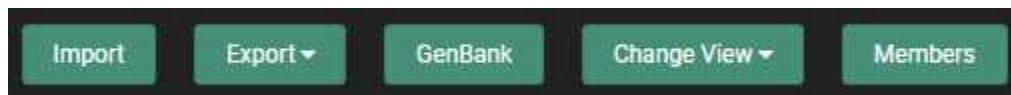
These buttons can be found at the top-right of your MycoMap project dashboard.

This takes you to the GenBank submission management screen. There are a number of tabs here. Click on the "Settings" tab. We will use the other ones later.

The GenBank upload default settings menu on a MycoMap dashboard.

These fields allow you to edit defaults that will appear on all of the records you submit to GenBank. For now, fill in the "Isolate Prefix" field. Whatever appears in this field will appear at the beginning of your GenBank sequence. I typically put my name and ONT. This allows me to quickly view my sequences in GenBank and to know the sequencing methodology that was employed.

Entering "S.D. Russell ONT"

Will produce a sequence title in GenBank similar to:

"Pluteus hongoi isolate S.D. Russell ONT iNaturalist # 17726745 small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence"

Secondarily, you will want to include some default notes for each record. I typically include the

project name, the type of sequencing used, as well as the flowcell and ligation kit for the run. Example:

"Mycoflora of Indiana; ONT MinION nanopore sequencing; Flongle 10.4.1; SQK-LSK114"

Hit save.

12   The next step is to begin examining the dashboard of your project and to begin examining individual sequence-observation associations. You should see something similar to this:

Individual lines on the dashboard of a MycoMap project.

For an overview of dashboard functionality, you can review this document online.

Begin by clicking on the BLAST "B" next to an individual record line. This will open up the BLAST results for the sequence records associated with the observations. If you have a popup blocker enabled, you will want to disable it for this website. If there are 3 associated consensus sequences, it will open up three new tabs with BLAST searches.

13   The BLAST page.

This protocol will not examine all of the eccentricities of a BLAST search, as this could really take up a whole book. Some brief annotations on the MycoBLAST and nanopore-specific considerations are found below.

For each BLAST search, there are actually two BLAST searches run - one based on NCBI data and another based on MycoMap data. You can think of the MycoMap data as a staging ground for future GenBank data.

Some things I look at first:

1. The summary fields at the top. If there is a single species name in the boxes for both NCBI and MycoMap, that is a good indication of what the ultimate identification will be.

2. The identity value, query coverage, and orientation of the sequence - in each BLAST area.

If you believe you have a good match and the name on iNaturalist is NOT correct, but the correct name is available on iNat:

Click on the title of the record on the dashboard. This will take you to the iNaturalist observation. Update the name appropriately on iNaturalist. If you are not able to independently get the record to reflect the proper species name (typically due to other competing identifications), then add the species name to the "Provisional Species Name" observational field.

Refresh the record on the MycoMap dashboard by clicking the green refresh icon seen below.

 The green refresh icon is located on the far right side of the dashboard.

If you believe you have a good match and the species name you would like to use is not available on iNaturalist, and the epithet has been validly published:

Request that an iNat curator include the name into the iNaturalist database. Email me.

If you believe you have a good match and the species name you would like to use is not available on iNaturalist, and the epithet has NOT been validly published:

Add the species name to the "Provisional Species Name" observational field. Put quotes around the epithet to indicate that it is a provisional name. Ex = Amanita "banningiana" or Amanita "sp-IN02"

If you would like to newly delimit a species-unit with a new provisional name:

1. Go to MycoMap -> Explore Species -> New Species Name

2. In the "Copy from" box, enter the species name you would like to use, to see if someone has already used it before.

3. If the name is open, enter it in the "Name" field at the top.

4. Select the proper rank from the dropdown. This will most commonly be "Species."

5. Select the closest relative to your species. It will pull all of the higher level classifications from that record. Ex - If you wanted to use "Russula "sp-IN102," then you could pull the higher classification from any other Russula species. Such as Russula vinacea.

6. Hit save.

7. Update the "Provisional Species Name" observational field on the iNaturalist observation with the newly created name. Also enter the name in the comments, with an additional note such as "DNA - ITS - Nanopore"

8. On the project dashboard, hit the green refresh box described above. It should refresh the record with your newly created name in the MycoMap project.

If you believe you have a good match and the name on iNaturalist is correct:

Clicking the checkmark once turns the circle green. Clicking it again will turn it red. This allows you to track the status of your analysis for a particular record from the dashboard, and allows you to filter for records that are finalized or that may need more work.

14   Prepare your record for GenBank upload.

Next to the BLAST "B," icon on your dashboard, there is the NCBI icon - a small blue double helix. Clicking this opens the GenBank validation screen.

Species Name: This will be pulled from the iNaturalist record. If you use a provisional name, the screen should automatically put the name into the right format for NCBI. Ex - Amanita sp. 'banningiana' or Amanita sp. 'IN01'

Collection date: Pulled from the iNaturalist/MO record.

Location: Edited from the geocoordinates of the iNat record to be in NCBI format. Beneath you will find the original text from the iNat record as well. Update as needed. Be sure to use the NCBI format if you make corrections:

Country: State, County, City, Location Name.

Lat/Long - Automatically pulled from the iNat record.

Isolate: This is the text that will appear in the title of your GenBank record, directly after the species name. You should have previously set some default text in the GenBank settings (previously outlined). For my records it is similar to: S.D. Russell ONT iNaturalist # 121781615

Specimen-voucher: This will often be blank unless you have already submitted the specimen to herbaria and have received back accession numbers and/or MyCoPortal numbers. You can always add this information in at a later date.

Sequence: If there are issues, this field will be flagged with a red X. If you have a green checkmark, you are ready to proceed.

Forward Primer Name: Should be populated with the information you entered in during the initial ZIP file import.

Reverse Primer Name: Should be populated with the information you entered in during the initial ZIP file import.

Collected by: Pulled from the iNaturalist record. If the user has entered their name in their iNaturalist profile, that will be used. If the observational field "Collector's name" has information, that will be here. Otherwise it will likely be in the format: iNaturalist.org User: stevilkinevil. Update this field as needed.

Identified by: Same as collected by. Update as required. I typically change this to my name for all records I validate.

Notes: There is a default for this in the settings that you should have previously updated. My notes typically look something like this: Mycoflora of Indiana; ONT MinION nanopore sequencing; Flongle 10.4.1; SQK-LSK114; iNaturalist.org #121782791

I will typically validate all of the information and then hit the master green checkbox at the top. This will put a green validation checkmark next to each aspect of the entry. Finally hit save.

If all of the entries have been validated for a record, then the NCBI icon on the individual record line will now have a green background.

****Note, if your record does not have the NCBI icon, it is likely that the sequence for a record is A.) has one of the quality check flags triggered [low quality, wrong specimen, etc]) or B.) the sequence is under the 200bp minimum requirement for GenBank Upload.

Issue: Multiple Sequences for an Observation

15 It is possible that a number of the resulting observations have more than one sequence associated with the record. I have seen up to 16 sequences attached to a record. This is due to a number of factors that are explained in the following steps of this section. Clicking the sequence "S" box on an observation allows you to see the sequence records that are associated with an observation. The image below shows two sequence records. One is low quality and with the wrong orientation, so it is easy to assess which is likely to be the best one to utilize from this screen alone. I will manually take a look at the BLAST results for each and every sequence that is associated with an observation.

Add Sequence ✕

S No GenBank link to the sequence	Import to iNat

Sequences

S H02-OSF2022.0268-iNat130056787-2   🕐 ❌

PROBLEM FLAGS
- ⊘ This sequence is a contaminant
- ⊘ This sequence is the wrong specimen
- ⊘ Low quality sequence
- ⊘ Wrong Orientation

S H02-OSF2022.0268-iNat130056787-1   🕐 ❌

PROBLEM FLAGS
- ⊘ This sequence is a contaminant
- ⊘ This sequence is the wrong specimen
- ✅ Low quality sequence
- ✅ Wrong Orientation

Blast Searches

	Run Blast Search
H02-OSF2022.0268-iNat130056787-1 - GW1YV1E2016	August 30
H02-OSF2022.0268-iNat130056787-2 - GVW0ZZAD013	August 30

16   **Problem: One sequence is in the wrong orientation**.

Occasionally, the pipeline will output one sequence in the right direction and one sequence in the "minus" direction (the reverse complement). This is likely to be the scenario if there are two sequences, where one is flagged as low quality, while the other shows no flags.

Solution: Check both sequences via the BLAST search to ensure that both match the target specimen. Be sure to double check the query coverage for each as well. If there is a target sequence with the right orientation, move the sequence with the wrong orientation to the trash, while maintaining the association with the sequence that is in the right "sense."

17   **Problem:** When checking the BLAST results, there is a **low query coverage for your sequence**.

Problem A: An adapter or A-tail sequence remains attached to the sequence.

This can be discovered using two different methods. First, visually examine the nucleotides of the sequence on the record page for the sequence. The color-coded nucleotides make it simple to identify an adapter. They will appear as a long

A long G-tain (the reverse-complement of an A-tail) is still attached to this sequence. There is also a short adapter sequence that remains. This sequence should be manually trimmed if there are no other target sequences, or moved into the ONT Trash if there are other associated sequence records of the target without the adapter.

Solution: Trim or discard the sequence.

Problem B: Your sequence is a chimera - a aggregate of two or more organisms into a single sequence.

This seems to happen more frequently with Illumina libraries, rather than nanopore, but it does occur occasionally. They are most easily recognized when you have a low query coverage combined with strange BLAST results. The results often have multiple unusual species in the results (sometimes combined with your target) or may appear as a contaminant with low query coverage. I have seen chimeras with up to three different sequences combined into a single ~600bp read. (Or even 3 within a 250-300bp Illumina read).

Example of a chimera with a protozoan.

Solution: Trim or discard the sequence.

18   **Problem: Non-target sequences are associated with the observation.**

The additional sequences could be explained by the bioinformatics pipeline being employed. Nanopore has an inherently high error rate. As such, we designate the maximum amount of errors that are allowed in the tags to make an attempt to associate with a given observation. This allows more of the sequences to pass the QC filtering process, but results in more analysis that needs to be done on the backend to clean out the non-target sequences. These non-target sequences typically have a low RiC and a normal query coverage.

Solution: Manually review the BLAST results.

Possible Solution 1: If all of the sequences have a low RiC and none of them match the target, this typically means that sequencing was unsuccessful for this specimen. Move non-target sequences into the ONT Trash, even if this removes all of the associated sequences.

Possible Solution 2: If some of your reads are common yeasts or fungal decomposers, particularly if they have a high RiC, it may be worth leaving the sequence associated with the observation and not moving it to the trash. Flag the sequence as a contaminant on the sequence record page. Building a database of these specimens may help to give us interesting ecological information for future use.

Possible Solution 3: If there is a high RiC, it is possible that the specimen was mixed up during extraction, database entry, or library prep. I will usually flag the record as "Wrong Specimen" and circle back to them at the end of the analysis process, to see if there are other specimens in a similar situation that they may have been switched with. If no obvious solution presents itself, it may also be worth going back to the original specimen to ensure that specimen matches the observation. If no solution can be found, remove the observation association from the sequence record and leave the sequence in the normal database category (ONT Sequences).

Upload your Sequences to GenBank

19   Following this protocol, you should now have many records validated on the NCBI screen. You can submit one large batch or multiple smaller batches. The choice is up to you. The time it takes is similar for either submission strategy.

Go to the GenBank Submission Portal. Register/Login.

20   On your MycoMap dashboard, click the green "GenBank" button at the top-right. The "My Verified Records" tab should list all of the sequences that are ready for submission. Select the individual records you want to submit or use the select all box at the top.

The GenBank Verified Records screen within a project on the MycoMap dashboard.

Hit the "Download" button at the bottom of the screen.

21 You should now see two newly generated files in your project - one called "Source Modifiers" with a "TSV extension and another called "Sequences" with a .fasta extension. Download both of these files to your computer.

The two files that will need to be uploaded to GenBank.

22 Within the GenBank Submission Portal, go to New Submission.
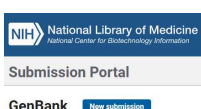
Submission Type:

What do your sequences contain?   rRNA or rRNA-ITS

What type of rRNA or rRNA-ITS?   Eukaryotic nuclear rRNA or rRNA-ITS

What do these eukaryotic nuclear rRNA or rRNA-ITS sequences contain?   contains rRNA-ITS region

Submission title - whatever you choose. Ex - ONT05-1

Submitter: Enter any affiliations you have. The person entered here will be the owner of this record.

Method

What methods were used to obtain these sequences? Other "Oxford Nanopore MinION; Flongle 10.4.1"

These sequences are: Assembled sequences

Assembly Program: NGSpeciesID

Version: v0.1.2.1

Sequences

Select a release date.

Upload the Sequences.fasta file that you downloaded from the MycoMap project.

Hit continue. It may take a few minutes to process the file, depending on the size.

Source Information

Do your sequence IDs represent one of these?   None of these

Which of these source modifiers would you like to use to uniquely identify each sequence in your submission? **Isolate**

Continue

Source Modifiers

Upload a tab-delimited table.

Select the Source Modifiers TSV file that you downloaded from the MycoMap project.

Continue

At this point you may see one or more warnings. Just continue. I have never needed to do anything at this step.

If you are using any new provisional names, you will need to fill out the Additional Organism Info screen. The new names should be populated. Just select "New Species" all the way down. No need to fill in any of the other fields.

References

Apply the authors and references as appropriate. For reference, I use something like:
Unpublished
"Mycoflora of Indiana - Nanopore Sequencing Run 05"

Finally review and submit. You are done with the first portion. Typically in a couple days, you will receive a notification that your sequences have been accepted. You will have one more process to do at that time.

23 **Submission of the Accession report to MycoMap**

Submitting your sequences to GenBank is only half of the process. The other half is to submit a file to MycoMap that links all of your new GenBank accessions back to the iNaturalist records.

Once you receive an email saying your submission has been accepted, go back to your GenBank Submission Portal dashboard. You should see the following. Download the AccessionReport.tsv file.

Go back to your MycoMap dashboard. Green "Genbank" button on the top right. "GenBank Confirmation" tab. Upload your AccessionReport.tsv file here. In seven days, the project will

refresh all of the GenBank records on your dash and make the appropriate linkages (Even though you requested immediate release, they are never available right away once you get the confirmation). All of your iNat records will now be associated with GenBank accessions on MycoMap.

24   Import data back to iNaturalist

Once your GenBank data has been made available and refreshed within your project, you will want to push all of your results back to iNaturalist. It is possible to do this for all verified records at once.

Import -> Refresh -> iNat observational fields -> Entire project
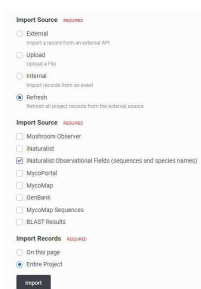
This will ensure all of the sequences, links to raw sequence data, links to BLAST results, GenBank accesions, etc are all back on all of the iNaturalist observations.

Pushing Data Back to iNat

25   Once all of your internal data analysis is complete and your GenBank data is associated with all of your observations, you will want to push all of information back to the original iNaturalist observation. This can be done with a single button click from the dashboard:

Import -> Refresh -> iNaturalist Observational Fields (sequences and species names)

The sequence, BLAST results, link to the raw data, and GenBank information will be imported into the original iNat record. Only records that have been validated with a green checkmark will have the sequence data uploaded to iNaturalist.

Observation Fields (6)

DNA Barcode ITS:
GTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTACTGAACG
AAATGGGTGGCAAGGCTGTCGCTGGCTCGAATGAGCATGTGCACGTCTTTT
GCTGCTTGCTTCATTCTCTTCTCCACCTGTGCACTCTTTGTAGACACTCGGG
ATGTGAGAGAGGTTAGCATTGATTGTTGACCTCTCTTGATATTGAAAAGTCT
GGGTGTTTATGTATTTTTTGACATACACGTTTGAATGTCTATAGAATGAAAAT
GTAGGCTTTTGTCAGCCTTTAAATGATAAAATACAACTTTCAACAACGGATC
TCTTGGCTCTCGCATCGATGAAGAACGCAGCGAAATGCGATAAGTAATGTG
AATTGCAGAATTCAGTGAATCATCGAATCTTTGAACGCATCTTGCGCTCCTT
GGTATTCCGAGGAGCATGCCTGTTTGAGTGTCATTAAATATCTCAAAAAGCT
TGTGCTTTTTTGGCACAGGAGTTTTGGACATTGGGAGTTGCCGGCTGCTGG
ATAACAGTGGTGGGCTCTTCTGAAAAGCATTAGTTGAGGAGCTTTGCACTC
TATTGGTGTGATAGATTATCTATGCCAGGAGACGCTTCATGATCCTCTGCCAT
CTTAACCGTCTTTATAAGACAATATGATAAACTTGACCTCAAATCAGGTAGG
ACTACCCGCTGAACTT

MycoMap BLAST Results:
https://mycomap.com/genetics/blast-search/c10-bc34-inat131020155-1-
r47119

Provisional Species Name:
Amanita "flavoconia-01"

Trace Files (Raw DNA Data):
https://mycomap.com/genetics/sequences/ont_sequences/c10-bc34-
inat131020155-1-r45879

Voucher Number(s):
BC34

Voucher Specimen Taken:
Yes

iNaturalist observational fields with sequence data that has been pushed back to the original record.