DNAMA Volunteer Sequence Validation Corps Guidebook v3.0

By Harte Singer (FUNDIS, Dikarya LLC, Cal State East Bay) with major contributions from Joshua Birkebak PhD (DNAMA) , Alden Dirks PhD (DNAMA), Jessica Williams (FUNDIS, OMDL), Scott Ostuni (FUNDIS, OMDL, Mycota Lab), Stephen Russell (Mycota Lab). **Partially supported by funds from the California Institute for Biodiversity (www.calalive.org)**

Built upon a foundation of community contributions and old fashioned trial and error.

Overview of Sequence Validation

The primary objective in "validating" a sequence is to determine if the sequence is actually derived from the intended organism *and* that it is biologically accurate. A secondary objective is to assign a taxonomic classification to that organism based on a synthesis of available information: matching to reliable records in DNA sequence databases, macromorphological and geographical information from an iNaturalist or Mushroom Observer record, and the input from members of the community at large.

This guide will provide you with all of the information you need to begin validating sequence records generated following Stephen Russell's Oxford Nanopore protocol.[1] Ultimately the choices you make will be based on your judgment and in many cases there is no "right" answer; you should feel empowered to make the least-wrong choice. Science is an iterative process that is driven by errors, blunders, and mistakes. As scientists, we must feel confident in our determinations but also demonstrate the humility and flexibility to be wrong. This program is designed to provide robust support for you to build your confidence—a distillation of the lessons learned from many years of errors, blunders, and mistakes. You should always feel comfortable reaching out via Slack or during our regular meetings to ask any questions or review any of your work. In the end, our work will be entirely in the public eye and subject to critique. It is this form of consensus that moves the institution of science forward as a whole. Please use this "live" version of the document to provide commentary
https://docs.google.com/document/d/1puazaKY2KJq5lW0GH6X_em2u8H8Mk_TyQpsjBoRRmPg/edit?usp=sharing

[1] Russell 2023, dx.doi.org/10.17504/protocols.io.36wgq7qykvk5/v3

Introduction to Cell and Molecular Biology

The fundamentals of sequence validation involve comparing DNA sequences and combining that with photographic evidence and personal intuition. One can learn to do this with little understanding of biology—just as one can learn to drive a car without knowing how an engine works or how to change a tire or the laws of motion that determine how fast we will stop when applying the brakes. This program is designed to not only teach you to do the validation but also to understand what you are doing and why you are doing it.

If you have never taken a college-level general biology for-majors class, **I strongly recommend** you read through all of the suggested material before beginning validation work. I have also provided this information in several "speed-run" style videos which will cover the core concepts of fungal biology, cell and molecular biology, and the DNA sequencing technologies that produce the sequences you are validating.

The written material comes from freely accessible textbooks from OpenStax.

Foundational Concepts

"Speed Run" video - *https://youtu.be/yaBw10hqFN0*

Eukaryotic Cells: *Focus on the Nucleus and the Ribosome* *https://openstax.org/books/biology-2e/pages/4-3-eukaryotic-cells*

The Cell Cycle: *What happens during mitosis?*
https://openstax.org/books/biology-2e/pages/10-2-the-cell-cycle

Enzymes: *What is an enzyme?*
https://openstax.org/books/biology-2e/pages/6-5-enzymes

Process of Meiosis: *Define haploid, diploid, recombination*
https://openstax.org/books/biology-2e/pages/11-1-the-process-of-meiosis

Sexual Reproduction: *Compare/contrast different strategies*
https://openstax.org/books/biology-2e/pages/11-2-sexual-reproduction

Genetics

"*Speed Run*" Videos
Part 1 https://youtu.be/PxsQZuEAa5Y
Part 2 https://youtu.be/mOhIRYrbUKg
Part 3 https://youtu.be/SZT-jXtL_v8

Characteristics and Traits: *Define phenotype, genotype, allele*
https://openstax.org/books/biology-2e/pages/12-2-characteristics-and-traits

DNA Structure and Sequencing: *Understand the structure of DNA*
https://openstax.org/books/biology-2e/pages/14-2-dna-structure-and-sequencing

The Genetic Code: *"The Central Dogma" transcription, translation -*
https://openstax.org/books/biology-2e/pages/15-1-the-genetic-code

RNA Processing: *Describe types of RNA and their function*
https://openstax.org/books/biology-2e/pages/15-4-rna-processing-in-eukaryotes

The Ribosome is the translation machinery: *Understand the parts of the ribosome*
https://rotel.pressbooks.pub/genetics/chapter/the-ribosome-is-the-translation-machinery/

Ribosomes and Protein Synthesis: *How do ribosomes work?*
https://openstax.org/books/biology-2e/pages/15-5-ribosomes-and-protein-synthesis

Types of Mutations: *Understand DNA mutations*
https://rotel.pressbooks.pub/genetics/chapter/part-i-types-of-mutations/

DNA Damage Causes Mutations: *Identify reasons for mutations*
https://rotel.pressbooks.pub/genetics/chapter/part-ii-dna-damage-causes-mutations/

Haplotypes: Understand intragenomic variation
https://rotel.pressbooks.pub/genetics/chapter/haplotypes/

FUNGAL DIVERSITY SURVEY

Fungal Biology

Entire Chapter: *Overview of fungal biology*
https://openstax.org/books/biology-2e/pages/24-introduction

DNA Barcoding, PCR, and Sequencing

Video: https://youtu.be/t6evKAYBHPE

Slides: DNA Barcoding, PCR, and Sequencing

Workflow

- ***Investigate sequences*** *associated with each record for contamination/chimerism/errors.*
- ***Ensure sequence matches expectations*** *based on photographs and other metadata. If not, flag and trash sequence.*
- ***Review MycoBlast results****. Conduct additional NCBI BLAST searches as needed.*
- ***Review the FASTQ files*** *for polymorphisms, haplotype variation, chimerism etc.*
- ***Determine identity or assign provisional name*****
- ***Add sequence, MycoMap BLAST URL, Sequencing Technology , and Provisional Species Name*** *as well as any notes, links, discussion to iNat or MO*
- ***Update spreadsheet*** *with pertinent information.*
- ***Check the "verified" button in MycoMap*** *and add notes as needed.*
- ***Refresh the page in MycoMap*** *to update the name in MycoMap*

Slack

<u>Once you have completed the assessment</u> and joined our group, our day-to-day communication will take place within the VSVC workspace on the Slack app. You can use it on your computer or use the app on your smartphone. Here are some videos to get you started https://slack.com/help/articles/360059976673-Slack-video-tutorials

When you join, please introduce yourself in the #welcome channel and share some things about yourself such as what clubs you are affiliated with, what your favorite mushrooms are, and your social media contact if you would like. You can get help with tricky sequences in the #troubleshooting channel, and ask for provisional name approval in the #name-request channel.

The Validation Tracker

Sequence Status

Good - The sequence is at least greater or equal to 30 RiC and does not appear to be a contaminant. Looking at the alignment, there are no odd indels or other potential artifacts that require further review. There is no reason to suspect the sequence is problematic.

Failed - The most obvious failure is a cluster of sequences with very low read counts, usually they are automatically placed into ONT trash. Any record that has one or more sequences where there is no sequence >8 RiC is considered "failed".

Wrong Specimen - In this case there are one or more sequences, but none of them are from the target specimen. *They are either from a different sample in the run or they are an environmental contaminant of another mushroom*. What we are communicating is that the sequence is an error, and did not actually come from the specimen being sequenced. We do not use this flag if we get mold or yeast as those are likely to actually come from the specimen. This is often a bioinformatics issue, but in some cases this may indicate a sample switch or physical contamination. Mark it as "Wrong Specimen" on MycoMap. Include putative ID of the contaminating sequence on the sheet. Do not export to iNaturalist. <u>Place the sequence in ONT trash but **do not use the red X** to delete it.</u>

Off Target - The sequence matches to a *mold*, *yeast*, *or known mycoparasite*. The sequence is not from the target mushroom, but rather represents an organism that was very likely on the surface or within the tissues of the mushroom pictured. If there is only one sequence, you must decide whether to import it to the iNat record or not. Use your judgement to determine the likelihood that this is a “real” sequence. If there are multiple sequences, and one includes the “host” or the pictured mushroom, message the observer on iNat and ask them to create a new record for the other organism recovered. Mark the sequence as “Contaminated” on MycoMap but do not put it in the trash. Transfer it to the new record if one is created.

No Sequence - There is no sequence for this record - *Make sure to check the trash for sequences before reporting this.*

RiC < 30 - If this sequence matches reference data, or other records with 100% adjusted score, then this is likely to be a good sequence. In all other cases, the sequence must be manually reviewed by alignment. There is no hard RiC cutoff, although I would be very cautious using sequences with <RiC 10. Your confidence will be dependent on the quality of the sequences in the fastq file.

Revise - The sequence is in need of revision, either because it has a few suspicious errors or because there are polymorphisms that can be identified. Use this flag to identify which sequences need revision, and remove the flag once the sequence has been revised.

Polymorphic - This sequence has significant verified polymorphism (around 1% or greater). Use this flag to identify these sequences for future analysis and tracking purposes.

Record Status

Validated - A name was given to this record and it is consistent with the sequence results. In rare cases, this sequence may not be good enough quality to add to iNat or GenBank, but we know that the specimen almost certainly matches the sequence. This could be due to a very low RiC (< 8) or in the case of Sanger sequencing, a very short (< 200 bp) sequence.

Flag - There is something wrong or uncertain with this record and it will need to be looked at. Use this to flag any record that should be examined further by folks on the team with more experience. You can use it to request review of your own sequence or to flag a teammate’s sequence. While our job is not to nitpick each other's work, we should not hesitate to identify when something is problematic so we can correct it before publication. We must all be comfortable with our work being scrutinized.

Bookmarked - This record is in progress and validation is incomplete due to needing to stop working or take a break. Use this flag to mark records to be completed later.

Stalled - This record is in progress and validation is incomplete pending further information. Use this flag if you have reached out to someone for help or context or are otherwise waiting for some information before completing the record.

Other Fields

Identification - This is where you make a determination about the identity of the specimen. It is your current ID for the record. If there is a contaminant, identify the contaminant here. Contaminant ID may be useful in tracking mycoparasites or misidentifications, or in troubleshooting confusing results.

New Name - A new temporary provisional name is needed for this record. There is a requisite observation field in iNaturalist "New Provisional Name Created" which accepts a date as the value, make sure to use that feature on iNaturalist in conjunction with this field on the sheet. All new names will need approval. Enter your proposed name in this field and request the name be added in the "name-request" Slack channel.

No Match > 99% - We use this to indicate if the BLAST results match > 99% to anything in NCBI GenBank and in MycoMap local results. While there is no % identity threshold for species delimitation - in some rare cases a single nucleotide separates species - we can get a rough estimate using 99% as that threshold. This metric will be used in our reporting to indicate how many records we sequenced that had not been sequenced prior.

Regional Novelty - is this the first *sequenced* record of a taxon in a particular geographic place? We are using an approximate 99% pairwise identity with at least 80% query coverage in most cases as a cutoff for this determination.

Notable - This is for tracking a number of different qualities.

a. Known Rare Taxon - This record is of a species known to be rare, *i.e.*, a FUNDIS rare challenge species or IUCN Red List species.
b. Parasitism - This record is evidence of parasitism of one fungus on another.
c. Not resolved with ITS - This taxon cannot be confidently identified using an ITS sequence alone.
d. Closest match >5% - This sequence is fairly distant from other taxa.
e. Cryptic Species This is a fairly subjective assessment that the specimen belongs to a taxon which is indistinguishable from other related taxa based on visual appearance alone.
f. Multiple Haplotypes This individual contains at least 2 distinct haplotype variants that you separated (phased) into 2 sequences
g. Sequence is rare At the time of analysis, there were <5 matching sequences between all of databases on MycoMap.

h. Other - Write in the notes anything else that is notable.

iNat checkbox - Mark this once the data has been published to iNat. Only publish sequences to iNaturalist if you are confident that they are both accurate and match the specimen depicted in the iNat record. There should only be one sequence associated with the record on MycoMap before doing this unless the other sequences are also good (***trash any bad sequences***). Add your name to the iNaturalist record in the “Sequence Validator” observation field.

Complete checkbox - Mark this when everything for a record has been completed. This is independent of the validation status. It is simply meant to track records that have been looked at and completed regardless of the outcome.

“Step By Step” approach

1. **Choose a sequence to work on from the spreadsheet.** Add your name to show others you are working on it. Open the associated MycoMap project in another tab. Use the iNat or Mushroom Observer number from the spreadsheet to locate the MycoMap sequence record. Open up the sequences tab and MycoBLAST results in a new tab for each sequence.

2. **Review the sequence page.** Open up the associated observation in a new tab. How many sequences are associated with the record? What are the read counts? Are there any problem flags on them from MycoMap (i.e. Low Quality)?

3. **Review the MycoBlast Results for each sequence.** Is it obviously a contaminant? Is it obviously chimeric? Does it have a really low read count? Place these in the *ONT trash* and continue with the most likely correct sequence. Make sure to add the appropriate flags “contaminant” and “low quality” to the record in MycoMap by marking the requisite checkboxes on the page. Remove the “low quality” flag from anything that is actually a good sequence. As you move through, take any bad sequences and add them to ONT Trash.

4. **Review the FASTQ files.** In general, I recommend reviewing the raw data for every record. Records below 30 RiC should *always* be examined using UGENE to determine the reliability of the sequence. Even sequences with high RiC numbers can have errors or be hiding polymorphic sites and significant haplotype variation. Once data is published to NCBI GenBank, it is much more time-consuming to fix, so when in doubt, review the FASTQ. As you become more experienced and familiar with specific taxa, you may choose to skip this step when you are confident that there will be no errors based on the BLAST results.

5. **Create a consensus sequence and phase haplotypes if needed.** If there is any polymorphism within a FASTQ file, create a consensus sequence showing the most information possible. That means applying IUPAC codes for any position where at least 25% of the bases are different from the dominant base (threshold ~75%). Any time there is length variation, include all of the nucleotides from all variants (no gaps). When possible, phase out the individual haplotype variants ***and*** create a consensus sequence. Make note of your edits on the sequence page, and when you post to iNaturalist so people know this is a manual consensus.

6. **Edit the sequence on the MycoMap sequence page.** Replace the sequence in MycoMap with your new consensus and adjust the number of reads used and append the sequence record to communicate that it was edited. You will need to be approved on MycoMap in order to have permission to do this. Finally, create a new MycoBlast search with the edited sequence and replace the MycoMap BLAST search link with the newly created link. ****For haplotype variants**** Each variant gets a separate MycoBlast search. The consensus of both haplotypes replaces the main sequence. There are fields in iNaturalist to put haplotypes and their respective MycoBlast searches.

7. **Is there a putative match?** Anything > 99.5% match to a TYPE sequence or any other reference quality sequence is a strong line of evidence to consider a match. 99.5%-100% matches to a large group of sequences with the same name that do not appear to be polyphyletic is another good line of evidence to support your choice. Always check on Index Fungorum to see where that taxon was described from and when. Sometimes many folks misapply a name and we may want to correct that when we have strong evidence to do so. Annotate sequences in the MycoMap BLAST results to use the correct name. Add the *sensu stricto* observation field to iNaturalist if the sequence is true to the taxon name.

8. ***If* there is an unresolvable problem** with the sequence (i.e. possible contaminant, artifact, or chimera, and there *are no other good sequences)*, flag it appropriately in the spreadsheet and mark the "verified" check mark button red in MycoMap**.**

9. **Conduct additional NCBI BLAST searches as needed** to potentially resolve the sequence, *i.e.*, searching against specific taxa, changing search parameters, or > 50 search results.

10. **Verify the morphology of the specimen.** Look for unique and identifying characters in the iNat/MO observation. Use ***other observations*** with trusted ID's or the representative photos of the taxon. If the species is listed on https://www.mykoweb.com/ or, www.mushroomexpert.com, check there. You can also sometimes find photos on GBIF.

11. **Make your best possible taxonomic determination.** If it is a provisional name, then report that in our spreadsheet as well as any records that helped you determine that. If it is a binomial, make sure to confirm that the name is current and accepted using MycoBank or Index Fungorum. Report all of your notes in the spreadsheet and in iNaturalist or Mushroom Observer.

Using ITS to establish species synonymy:

ITS barcoding is a powerful tool to aid in species identification and delimitation of novel taxa. It is not a "silver bullet" and should not be used to dismantle established taxonomy without additional lines of evidence. If a sequence matches 100% with more than one type sequence, we can suspect that they are essentially the same taxon. However in some taxa, morphologically and biologically distinct species differ by only a single nucleotide, and other taxa may be impossible to separate into distinct species using ITS barcoding. A few genera are known to contain species with intraspecific variation as high as ~ 5% ITS pairwise similarity. Some taxa appear to exist as a spectrum of closely related interbreeding populations and are best left as a "species complex" rather than attempting to separate them into distinct taxa. One should always consult supporting literature that provides strong evidence, *e.g.*., morphological and microscopic characters, ecology, and geographic distribution, before deciding to assume synonymy between two published names. Our goal is *not* to rewrite taxonomy, but we must report our conclusions should they suggest that current taxonomy is incorrect. I would recommend consulting with the community before deciding to synonymize published names that are currently recognized as different.

Here is a running list of groups for which ITS is especially problematic
DNA problem groups
In some cases you can use a combination of evidence, morphological, geographic, host, and DNA sequence to resolve some of these to species, in other cases it is best left at the genus level.

<u>Provisional naming philosophy and guidance:</u>

Please review the following guidance on provisional naming from OMDL
https://namyco.org/wp-content/uploads/2024/02/OhioDNALab.pdf

Ensure you are using the best provisional name *and it is not already taken*. The best way to do this currently is via MycoMap.com (instructions in the slides). You can also see a list of all provisional names on iNaturalist here:
https://docs.google.com/spreadsheets/d/1bA4JofqR-ltDLUo0EgTFgtF84kAtEBMXG3rMFW6bAQQ/edit?usp=sharing

Make sure to check MycoMap before suggesting any provisional name. If there are multiple provisional names for a taxon —*i.e.*, one from Steve Russell and one from Danny Miller—we will default to Danny Miller in most cases. We will always choose a provisional name that is pre-existing for something, even if it is not in the format that Steve or Danny uses. Rod Tulloss has many provisional names for *Amanita*, so if we see *Amanita* sp. 'cochiseana' in genbank, that is what we will use. Until you are comfortable, provisional names must be approved using the "name-request" Slack channel.

When identifying your provisionally-named record on iNaturalist, find the closest taxonomic rank that encompasses this record cleanly. If the one you are analyzing comes out close to a known taxon, identify it as such on iNaturalist. So for example if I am looking at a *Pseudohydnum*, and it looks macroscopically like *Pseudohydnum gelatinosum* and genetically it is closely related to other records of *Pseudohydnum gelatinosum*, then it should be ID'd as such in iNaturalist. In the ecosystem of iNaturalist, this makes sense since the vast majority of records are uncurated and unsupported by any meaningful evidence. By grouping things based on morphological similarity and genetic closeness, it allows for the amateur to better use iNaturalist in a field guide capacity, and it allows the analyst to more easily separate out discrete groups of data. It is not a very "clean" way of doing things and is by definition incorrect, but in this context, it becomes the most functional way to categorize things in the absence of a proper provisional taxonomy framework or the ability to designate *affinis*. <u>Note that this is a **stark departure** from past guidance which was to place something like this above the species rank.</u>

Using the *sensu stricto* observation field:

FUNGAL DIVERSITY SURVEY

In order to curate taxa that have strong evidence for being *sensu stricto*, we have created the requisite observation field. Please use this field liberally! It is a taxon lookup field, so you can enter part of a name and it will autofill the rest, or you can even search with a common name. This field allows us to curate records that we strongly suspect to be "the real thing". A match to a type sequence is the best evidence for this, however folks with expertise in a particular taxon need not rely on DNA data to use this field. It is an indicator of strong confidence and it allows us to override the community taxon. It will soon be implemented in MycoMap and allow us to clean published names out of the Provisional Species Name observation field.

Formatting notes:

We recently changed our formatting to match NCBI format. The two acceptable formats are:
Genus sp. 'stateNumber' (Mycena sp. 'CA01')
Genus sp. 'epithet-Number' (Mycena sp. 'maculata-CA01')

Correctly formatted MycoMap provisional names:
Pseudohydnum sp. 'gelatinosum-PNW01'
Agaricus sp. 'PNW09'
Ramaria sp. 'CA02'
Other formats you may encounter :
Mycena "sp-CA01"
Ramaria PNW18
Pseudohydnum "gelatinosum-PNW01"

12. **Import your data to iNaturalist** using the import to iNat button in MycoMap and copy over any relevant notes and matching observations. Verify that the data was added correctly. Report ONT Nanopore in the "Sequencing technology" observation field in iNaturalist. Enter the provisional name in the "Provisional Species Name" observation field if you have one.

Examples of validations

*Video 1:*https://youtu.be/00yrupqoEjM (Harte Singer: FUNDIS*)*
Video 2: https://youtu.be/Vs_yIHRfLOM (Jessica Williams: FUNDIS, OMDL)
Video 3: https://youtu.be/t_X2rKczSIM (Jessica Williams: FUNDIS, OMDL*)*
More video streams available at: https://www.youtube.com/@lothlin/streams

Guides for software and platforms

Using UGENE

Note: Older videos depict the need to search for "AGGATCATTA" in order to determine which reads are in the forward orientation and which reads are in the reverse orientation. This only applies to older sequences; our current bioinformatics pipeline automatically puts all reads in the forward orientation. These older videos are still very helpful and you will need to recognize when the data is a mix of forward and reverse reads.

Video(older): https://youtu.be/aUo1eqQyeoE (Jessica Williams, FUNDIS, OMDL)
Video(older) - common error :https://youtu.be/fj-96uwgqNU (Jessica Williams, FUNDIS OMDL)
Video: https://youtu.be/QlSDkNuRZKU (Harte, FUNDIS)

Haplotype Phasing in UGENE:

Slideshow: Haplotype Phasing in UGENE

Video: https://youtu.be/UrkFSnN9Ovc (Harte, FUNDIS)

Using Mycomap

Create an account on https://mycomap.com and email Harte or Steve to approve it.

Slides: *MycoMap for sequence validation*
Video: https://youtu.be/UwzxDj-JA14

Using NCBI BLAST

Video:https://youtu.be/eSMJWsPHyTI

Using iNaturalist

Video: https://youtu.be/UbrRYgja94M

Troubleshooting

Common challenges with ONT nanopore data.

Homopolymer repeats - Strings of the same nucleotide a.k.a homopolymers are the most common source of error with nanopore sequencing. The current across the pore that is translated into the DNA code only changes when different nucleotides pass through the pore. The raw signal from a sequence that is 6 of the same nucleotide will appear very similar to one that is 7 or 5 nucleotides in length, e.g. AAAAAA may be indistinguishable from AAAAA or AAAAAAA. Less often, strings of repeats cause a similar issue, e.g. ATATAT may be indistinguishable from ATAT or ATATATAT. These repeats also cause some issues with DNA replication machinery and sequencing-by-synthesis methods. When analyzing a sequence, we generally do not consider differences in these repeats to be phylogenetically significant. More often than not, the consensus sequence will report a longer sequence than is represented in the raw data. To manually annotate these regions, look for the length that appears most often (mode) and use that as a starting point to estimate the actual length. Sometimes these errors are the result of biological variation, but currently it is not possible to determine that. Use your judgment to make the best choice and ask for help if you are not sure.

Polymorphisms - Often, ITS sequences will have intragenomic variation - variation within the genome of a single organism. The automated consensus-building software does not resolve this. If you notice that your sequence differs by only 1 or 2 base pairs from other sequences, it is a good idea to align all of the reads in UGENE and look to see if there is variation at this position by copying and pasting part of the leading or trailing sequence from the BLAST results and then searching your reads for that motif. Likewise if the BLAST results show an ambiguity code, you should absolutely manually align the reads in UGENE to examine that position for polymorphy. If there is variation in the data, confirm that the variation is not due to poor alignment or frame-shifting. Report variation using IUPAC ambiguity codes. The most common will be Y (C/T) or R (A/G). In reality these are both the result of the same mutation - see the genetics section at the beginning of this document for more information.

Haplotype variation- An extension of simple polymorphism, you may see that each polymorphic site co-occurs with other polymorphisms. When there are several polymorphisms that co-occur in one sequence relative to the other sequence, this represents a unique haplotype. This term typically refers to looking at single-copy genes in diploid organisms where each haplotype is derived from each parent. In the case of rDNA, there can be dozens or even over 100 copies in the genome, and it is not unusual to see several distinct haplotypes. A very loose rule of thumb is: IF >1% of the sequence is polymorphic OR there is significant length variation AND this variation is separable into distinct sequences, THEN phase the haplotypes into distinct sequences. Add the longest-length consensus with IUPAC codes to the DNA Barcode ITS field on iNaturalist, and then add each phased haplotype to Haplotype A, Haplotype B, and if there is a third, Haplotype C fields on iNaturalist. Each field has its own associated MycoMap BLAST field.

Chimerism - The automatic consensus-building software will occasionally incorporate a chimeric read - a “Frankenstein” sequence made up of 2 different sequences, or one sequence that has had a portion duplicated. When this happens, the consensus will be longer than normal and will be biologically inaccurate. The biggest clue that you have a chimera is that the query coverage in the BLAST results will be <90%. Most often it is <80%. Just because there is low query coverage does not mean that you have a chimera, but 99.9% of chimeric sequences will have low query coverage. Another clue is that the sequence is unusually long. Most fungi will have an ITS sequence between 550 and 700 base pairs in length. Boletales and Cantherellaceae are strong exceptions, and some ascomycetes will have inserts at the beginning of sequences that make them very long (group 1 introns). If you see a sequence that is >700bp in length with a query coverage lower than 90% you should be suspicious. Make sure to run a new NCBI BLAST search against at least 500 records and filter results by query coverage to make sure there are no hidden sequences that span the full length of your sequence before taking the time to make a new consensus. When you take the reads into UGENE and align them, the hallmark of a chimera is that one of the reads (rarely 2 or 3) will be longer than the other reads. The chimeric portion can be located anywhere in the sequence but most often it is near one of the ends. Remove the offending sequence(s) from the pool of reads and make a new consensus.

"Nonsense" chimerism - Very rarely, the consensus building software will combine sequences from different organisms into one consensus. This happens when the pool of sequences is a mix of different taxa and happens most often when there is low RIC or when there are many similar taxa in a single run. It is very easy to spot these with a sequence alignment in UGENE. You will see groups of sequences that are very distinctly different from each other. A BLAST search of each group will show different organisms, be careful not to conflate this with haplotype variation. With a bit of work, it may be possible to rescue the good sequence from the hodge-podge, but it may be better to simply request the sample be repeated.

If you encounter problematic sequences and require help, please report them in the #troubleshooting slack channel so we can solve them.

Additional Resources

Glossary of important terms:
https://docs.google.com/spreadsheets/d/11PScedKAFzEih_shOKoZJsQO4pbfd_WWzw_5h0jQ9yE/edit?usp=sharing

Phylogenetic tree of the Agaricales - https://agaric.us/

Index Fungorum (IF) - https://www.indexfungorum.org/ source for taxonomy including current accepted names

MycoBank - https://www.mycobank.org/ another taxonomy website, parallels with IF and a good place to start when looking through names.

MykoWeb - https://www.mykoweb.com/ Mike Woods' in-depth website focused primarily on California mushrooms. Contains a wealth of images and species descriptions to help match specimens to sequences.

Danny Miller's website http://alpental.com/psms/ddd/ - lots of sequence-based mushroom identification information. Danny has spearheaded sequencing of PNW mushrooms.

Mushroom Observer - https://mushroomobserver.org/ like iNaturalist but only for mushrooms

Phylogenetic tree builder - https://www.phylogeny.fr/ Easy tool for building phylogenetic trees.

Definitions of TYPE names
-https://www.floridamuseum.ufl.edu/herbarium/collections/type/about/

UNITE - https://unite.ut.ee/ curated fungal molecular database. The dataset that is most often used to generate metagenomics analysis for fungal OTU's. Not very intuitive to navigate, but the sequence data is mirrored in MycoMap's local database.

List of ambiguity codes for for nucleotides
https://www.bioinformatics.org/sms/iupac.html