**Analyzing DNA sequences**

This tutorial will describe **how to use DNA** as a tool to do the research necessary to **determine which species of mushrooms** (of the taxonomic group you are interested in) **are present in your geographical/ecological area**, so you can collect descriptions and photographs of verified bona fide collections of each species. DNA is a relatively recent tool that can help you in your research. While many papers have been written about how to use traditional methods like microscopy, chemical tests and cultivation to identify mushrooms, DNA is so new that there is little documentation about how to analyze DNA sequences.

Throughout this paper, we will be answering the example question: **Does *Cortinarius violaceus* grow in the PNW?**

Being able to confidently figure out what species a collection is once you have a DNA sequence requires these important things:

1. You need to sequence your own collections or **find other people’s sequences** from your geographical and taxonomical area of interest that you wish to identify.
2. You need **pre-existing reliable, good quality sequences of the species** that your mushroom might be, in the same region of DNA, to compare with.
3. You need to **know how to compare 2 sequences** to see if they might represent the same species or not.

For instance, if you have a sequence of *Cortinarius violaceus* that you are sure represents that species, and you compare it to your local sequence and determine that they are probably the same species, then you have identified with some confidence that *Cortinarius violaceus* is present in your area.

If, as often happens, you compare your sequence to known sequences and determine yours might not be the same, you may have a new species on your hands. You need to figure out if you do or if you don’t, and if you do, help describe it and give it its own name. More and more, DNA research is helping to discover that a mushroom that has been assumed for a hundred years to grow somewhere may not be what everybody thinks it is.

Unfortunately, none of those steps are straightforward to accomplish. If your mushroom is said to match 100% with a sequence labeled *Cortinarius violaceus* in GenBank, the odds that your mushroom is *Cortinarius violaceus* may still be less than 50%, as for some groups, fewer than 50% of GenBank sequences are labeled correctly. And if your mushroom only matches 70% (extremely poorly) to a sequence labeled *Cortinarius violaceus*, it might still actually be that species if, again, that other sequence was labeled wrong.

This document will cover the best practices I know of for performing these three tasks, and where the method is inefficient and unreliable (which will happen a lot because this is such a new technology) I will discuss what improvements could simplify and improve the techniques.

NOTE: DNA will never settle some arguments. Do you think of a group of collections as all one species or split them into 5 species? That is, and may always be, a matter of opinion. DNA won’t help you decide at what branch to declare a genus, family or species. It can only tell you how related two organisms are and let you use your judgement about where to draw the boundary lines.

**1. Finding sequences of interest that other people have made**

**ID is hard**: Since identification of a mushroom, even after you have a DNA sequence, is quite difficult, you do not want to assume that other people’s sequences were identified correctly. Many people who get a DNA sequence still incorrectly ID their mushroom. For some groups, <50% of mushrooms with DNA sequences are identified correctly. If there is a sequence that somebody else made in a taxonomic group you are interested in from a geographical group you care about, you need to double check their work and identify their mushroom yourself before taking that as evidence that a certain species was found in a certain place.

**When to sequence your mushroom**: Perhaps, before you sequence a mushroom, you want to know if somebody else in the same area has already sequenced the same species so you can save the time and expense. Even so, you might want to sequence yours anyway. If the species is hard to identify, they might have identified it wrong. If theirs was from another ecological region of the planet (or an unknown region), geographic distance alone means it might not be the same mushroom. In both cases, I would sequence my mushroom anyway.

**When not to sequence your mushroom**: If an easy to identify mushroom like *Cortinarius violaceus* was collected near me and already sequenced by somebody else, I will not necessarily bother to collect, dry and sequence one myself. I will study the sequence that has already been made. (This is only true if I know that there is no reason to believe there is more than one cryptic species here. Surprises happen all the time when people sequence things that they think are unnecessary, so you never know). Also, if a species is rare and I’m having a lot of trouble finding it, I will search to see if somebody else found and sequenced one near me that I can study so I’m not dependent on finding it myself.

**Collect FASTA files**

Collect all of yours and others’ sequences of interest in FASTA files, with the extension .fasta.

* I keep a separate file for each genus, and sometimes a separate file for each section of a genus if it is a large genus (like *Cortinarius* or *Russula*). Having more than one genus per file will be too unwieldy (as you will see in my “Making Trees” tutorial, working with more than 100 sequences at a time can be very slow).
* Put the sequences in the genus file that you think the mushroom really belongs in, ignoring what genus the mushroom is actually in. For instance, *Mycena oregonensis* belongs in its own genus and in its own FASTA file.

\*\*\* I NEED A TUTORIAL FOR FIGURING OUT THE GENUS SOMETHING REALLY BELONGS IN – ANALYZING TREES \*\*\*

**GenBank** - <https://www.ncbi.nlm.nih.gov/nuccore> - the universal repository of worldwide sequences

Unfortunately, there is no easy way to find sequences in GenBank from a certain geographical area, but I have a method that gives fairly good results. For the Pacific Northwest, UW in Seattle uses accession numbers that start with WTU (the 175 year-old original name of UW, Washington Territorial University). U of O in Portland uses OSC, British Columbia herbariums uses accession numbers starting with UBC or DAVFP. Perhaps the name of the state or province is mentioned in the descriptions.

Finding all the sequences matching those criteria **will find too many sequences**. Just because a mushroom was deposited at UW does not mean it was found in the PNW, it just means it might have been. A state might be mentioned in the description for reasons other than where the mushroom was found. And the location is often not mentioned, so this method **will miss many PNW mushrooms as well**.

Plus, GenBank has sequences of all gene regions, not just the region you are interested in (in my case, usually ITS). ITS sequences will have ITS at the beginning of a word in the sequence description, or they will spell out “internal transcribed spacer” so I only search for results that match with “ITS\* or spacer”.

Not all sequences in GenBank are labeled with the correct genus. Sometimes it’s just an error, but sometimes there has been more than one genus attached to a species over time and you never know which they chose, for instance *Alnicola* or *Naucoria*. Also, a species like *Gymnopus peronatus* has gone by *Gymnopus, Marasmiellus, Collybiopsis* and *Collybia*. The epithet ending has changed too, from *Gymnopus peronatus* to *Collybiopsis peronata*, so all you know about that species epithet is that it starts with “peronat”.

You must type a genus or species completely and exactly, or use the “\*” wild card character; it has no tolerance for a letter out of place.

Here is an example of a complicated query I type into the search box, to attempt to find species that may be listed under different genera, may have multiple spellings for the ending of the epithet, are ITS sequences only, and are from the PNW (British Columbia, Washington, Oregon and Idaho).

**(Gymnopus or Marasmiellus or Collybiopsis or Collybia) peronat\* (its\* or spacer) (british or ubc\* or DAVFP\* or washington or wtu\* or oregon or osc\* or idaho)**

For this paper, I want to see if somebody has already collected *Cortinarius violaceus* in my geographical area so I execute the following query:

**Cortinarius violaceus (its\* or spacer) (british or ubc\* or DAVFP\* or washington or wtu\* or oregon or osc\* or idaho)**

As of this writing, 14 results came up. Click on “Summary” and change that to “GenBank”. Now scroll down and read them all to see which ones are actually from the PNW. Some definitively state they were collected in WA, OR or ID. Some do not say, so you would have to ask the authors. It’s probably safest to ignore a sequence if you’re not sure where it’s from.

*Tip and Trick*: Mushrooms with SMI numbers were collected in Smithers, northern British Columbia, even though it might not say anything more precise than “Canada” in the description.

I conclude that enough people have collected this mushroom in the PNW already that I will not bother to go out looking for one myself. I will use the dozen or so collections that I can verify were from the PNW in my analysis in step 3.

**Collect the GenBank FASTA sequences**

Where you changed “Summary” to “GenBank”, now change it to “FASTA (text)”. Copy and paste the sequences that were from your geographical area of interest into your personal fasta file for that genus or section. This step can be a little annoying as you only want the ones you verified were from your area. If that number is small, you can get the sequence data one at a time by clicking on “FASTA” underneath the sequence name when you are reading the “GenBank” description.

**Fix the GenBank sequence names**

The sequence is probably called something long and cryptic. Your FASTA file might look like this:

>KJ920058.1 Cortinarius violaceus isolate RHP30078 voucher TENN:030078 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence

TGTTGTCGCTGCCCTTTCGGGGGATTTTGCACG…etc.

* I change the long ugly string to make it clear the mushroom looks like Cortinarius violaceus, but that hasn’t been proven, by using “cf” (compare with).
* I put the state/province it was found in.
* I keep the GenBank number for future reference, but I remove the unnecessary “.1” at the end.
* If the mushroom can be found on iNat or MO, I will include those numbers. Now my FASTA file looks like this:

>KJ920058 Cortinarius cf violaceus WA

TGTTGTCGCTGCCCTTTCGGGGGATTTTGCACG…etc.

**GenBank accession numbers**

You can reference a sequence by only its accession number inside a FASTA file if you want, if you are only ever going to use BLAST to analyze your sequences, but I don’t recommend it. The sequence name will be the long, ugly cryptic string and looking at it in your file or in a tree you’ll have no idea what it is or why it’s there. Plus, when you make a tree using MEGA instead of BLAST, it won’t work. It also requires you to put all the accession numbers at the beginning of the file, and once you have an actual sequence with a header character “>” in the file, you can no longer reference GenBank accession numbers. (For more information see chapter 2 of “How to make genetic trees”).

**GenBank has major flaws**: When the technology was new they never realized how important it would be to make sure the **location of each collection was discoverable** both by humans and by automated computer algorithms so they could produce a map. There is also no way to search for only **sequences deposited only after a certain date** so you can see only the ones you haven’t seen yet (wouldn’t it be nice to check back every year or so to see the new ones)? You’ll have to look at the dates yourself. Also, you only get somebody’s interpretation of what the sequence is, **you cannot get the original chromatograms** to double check if a sequence is corrupt or not, so you cannot correct any errors somebody made.

**iNaturalist.org**

More and more people are sequencing their collections and putting the results up on iNaturalist, which has a great system for searching for mushrooms in a particular geographical area, and has **accompanying photos and notes**, which are not available in GenBank. **Rarely, people are putting DNA sequences in the observations** you can look for, hopefully this will become more popular over time.

**You can easily search** for mushrooms of a **particular genus or family** etc. **found in a particular state or province**. You can then export those results into a spreadsheet that will include the DNA sequence, if they have included their sequence in a field called “DNA barcode ITS” (this is hopefully becoming the standard place to put the sequence). Then you can sort the spreadsheet on that column and see all the observations with sequences.

Let’s say you look at the sequences available on iNaturalist right now, and then you want to come back in a month and see the new ones that have been added since today. Unfortunately, as with GenBank, there is **no good way to only see observations that have had DNA added after a particular date** (the last time you looked) which means that you will see old results you have already seen. You can view only observations that were made after a particular date, but obviously (almost) nobody has the DNA results on the same day they collect the mushrooms, so we need a way to only see observations that somebody has modified (specifically by adding the DNA field) after a certain date.

1. Put your genus of interest into the search box (e.g. “Amanita”, or perhaps “Fungi” if you want all mushrooms) and click “View Observations” in the popup.
2. Click on “Filters” on the top right. Then “More Filters”. Set the “Place” to the country, state, province or county you are interested in. I sometimes do this entire process several times, once for each state or province (BC, WA, OR, ID). Sometimes I filter by “Pacific Northwest/Alaska” to get the entire region at once, but that includes parts of California, northern BC and Alaska, which are not my ecological area and not as much of interest to me. You can also add any other search criteria you are interested in.
3. Click “Download” in the bottom right of the Filters box.
4. In section 3, Choose Columns, I check the following as they provide the information I might want: id, observed\_on, user\_login, created\_at, url, description, place\_guess, latitude, longitude, positional\_accuracy, place\_state\_name, scientific\_name, field:collection #, field:dna barcode its. The final one is the most important. If for some reason “dna barcode its” isn’t offered, this isn’t going to work, and it will only work if that is the field they put their ITS DNA sequence in. Let’s hope they did.
5. Now “Create Export” and wait for it to be done or set yourself up to be notified. Download the spreadsheet and sort by the “DNA barcode ITS” column to see what sequences are available.

If I do this export for *Cortinarius violaceus* in “Pacific Northwest/Alaksa”, I download information for 725 observations, but only two Oregon collections have a sequence in the “DNA barcode ITS” column. They are MH266404 and MK949107. I will use these two sequences in my analysis in step 3 as well.

**Fix iNaturalist’s broken FASTA format**

* That DNA sequence you get from iNaturalist is not in proper FASTA format, like sequences from GenBank are. They might be missing the header line and only contain letters like “ACTGA”. They might look like they have a header line, but it is broken, without a carriage return/line feed after the header and before the DNA letters, so you have to fix it.
* You will want to put a meaningful header that includes the GenBank accession number (if known) the iNat # and the purported species name (with “cf” meaning it looks like it, but that hasn’t been proven yet) and also the state/province of the observation. So even though the spreadsheet shows this:

>MH266404.1 Cortinarius violaceus voucher JLF5959

GAAGTAAAAGTCGTA…etc.

* I will change it to this: (being very careful to press enter after I type iNat79820231 to make sure that the DNA letters are at the beginning of a new line).

>MH266404 Cortinarius cf violaceus OR iNat79820231

GAAGTAAAAGTCGTA…etc.

* I got the iNat number from the spreadsheet. More often than not, the spreadsheet will just contain raw letters and you have to put the “>” header line in yourself anyway.

**MushroomObserver.org**

MO’s tools for only finding mushrooms from a particular geographic area are not as good, and the ability to download sequence data is also not as straightforward. This might change in the future. I certainly have found MO observations of interest that have sequences attached and downloaded them one at a time.

**2. How to get reliable sequences of a species**

Step 1 is done: you now have local sequences (your own or somebody else’s) that you want to identify with confidence. Now for step 2, to find reliable sequences of that mushroom to compare yours with to see if it really is what you think it is. Remember, every mushroom sequence that we found in step 1 we labeled with “cf” meaning we didn’t trust the identity, no matter what. So the trick is, how do you find sequences you trust?

**2a. What if you have no idea what your sequenced mushroom is?**

Often you don’t know, that’s why you collected it. Or maybe you saw something cool on iNaturalist and got its sequence because it’s a mystery you want to solve.

You can get a general idea using BLAST. It will tell you some possibilities. Visit BLAST at <https://blast.ncbi.nlm.nih.gov/Blast.cgi> and click on “Nucleotide BLAST”.

* You want the default page with only one box to paste sequences into. If the page has both a Query box and a Subject box, un-click “Align two or more sequences” so you only have a Query box.
* Paste the sequence or its GenBank accession number (if it has one) into the Query box. At the bottom of the page, make sure under “Algorithm parameters” that “Max target sequences” is 100 and check “Show results in a new window”. Then click “BLAST”. For the *Cortinarius violaceus* from OR, I type “MH266404” into the top box and I get the image shown below.
* The list is sorted by “E value” (the computer’s guess at putting the most similar sequences on top of the list, but it often guesses wrong, due to an assumption they could easily fix (see the tip and trick below). You may want to click on “Per Ident” to sort on the percentage that the two are identical in overlapping regions where it could find similarity. If there are regions that were so dissimilar it could not align them, the percent identical will be way lower than shown, so there is no way to sort by actual percentage identical, the computer isn’t that smart. You have to examine the sequence comparisons yourself. It’s hard to say which way of sorting will put the actual most similar sequences on top, where you want them. You may have to try sorting both ways.
* If the “Max Score” and “Total Score” are not the same, then the beginning matched and the end matched, but the part in between didn’t. That means there are significant differences between the sequences, and even though one part may have matched at 100%, there are other parts that didn’t match at all. You may still want to look at the differences by hand, but there‘s probably no way those two sequences represent the same species, even with 100% “Per Ident”.

Graphical user interface, text, application, email

Description automatically generated

* For the ITS region of DNA (which I am using) anything 99.5% or higher might be the same species as your mushroom. Ours shows many matches that high with *Cortinarius violaceus*, so it is certainly worth pursuing the theory that our mushroom may indeed be *Cortinarius violaceus*.
* Anything >97% just possibly could be your species, so make a note to get reliable sequences of all of the names that match that high as well, especially if nothing matched close to 100%
* Anything >90% may be a different species in the same genus. If all of the matches >90% are the same genus, you may at least have figured out the genus of your mushroom.
* Anything >80% might be in the same family?
* If you are dissatisfied with the comparison, you can try comparing with more than 100 sequences, because BLAST is not actually showing you the 100 closest sequences, even though that is what it claims. Go back to the tab with the previous page and Change “Max target sequences” to 1000 (or even 5000 if you’re feeling patient) and when the results eventually come up in a second window, click on the column “Per. Ident” to sort by percent identical. You’ll see many more comparisons that way and get a better idea.

*Tip and Trick*: If you’re curious, here’s why **the top 100 matches that it gives you aren’t really the top 100 matches to your mushroom**. (You may have to ask for the top 5,000 matches to get the top 100 included in the results, and even then, some of the closest matches might still be missing). Why is that? Let’s say you sequenced all of ITS (ITS1, 5.8S and ITS2) which typically is successfully done. Some old, degraded or contaminated collections are difficult to get DNA out of. GenBank is full of sequences where only ITS1 or ITS2 alone could be retrieved. When BLAST compares your long sequence to one of GenBank’s shorter ITS1 only sequences, it notices that half of your longer sequence could not be matched up with the shorter sequence, because the data is missing, and it assumes that it would not have matched (that the DNA would have been entirely different). So if your sequence is identical to the GenBank one, it will think 50% of its length matched 100%, but the other 50% matched at 0%, so it will think that your two mushrooms are 50% different, when they are the same. A much wiser algorithm would assume all the unknown comparisons were good matches, or better yet, assume the same ratio of matching in other genes as you found in the gene you did match. For instance, if ITS1 matches at 97%, I’ll bet you ITS2 was going to match about 97% too, that’s just good math.

The problem: a shorter sequence than yours will always appear to be a poor match, and way, way down a list sorted by “E value”. If you sort by “Per Ident”, the shorter sequences that match 100% in their overlapping regions will show up on top, where you want them, but only if they were in the top 100 when sorted by “E value”. Often, a perfectly matching short sequence will be about 5,000th down the list when sorted by “E value”, so unless you asked for the top 5,000 matches, when you sort by “Per Ident”, the perfect match you were looking for won’t come to the top.

The workaround: **If it’s important you find matches, split your sequence into two parts, just ITS1 and just ITS2**, and BLAST them separately. You will get the actual top 100 matches in your list.

The real solution: BLAST could either 1) assume that areas that can’t be matched are good matches instead of bad matches or 2) when you sort by “Per Ident”, re-query GenBank for the actual top 100 “Per Ident” matches instead of simply re-ordering the top 100 “E value” matches.

**2b. Are there reliable sequences already available of the species you are interested in?**

**I’m making a list**

You can’t trust most internet sources, but maybe you can get a sequence personally from someone you trust, who can vouch for its authenticity.

I am spending a lot of time trying to get reliable ITS sequences of every Pacific Northwest mushroom (and others of interest) and giving them to everybody so they can compare their mushroom sequences to my list and get results more reliable than comparing their sequence to GenBank using BLAST.

* Visit my DNA website: [www.alpental.com/psms/ddd](http://www.alpental.com/psms/ddd). Search for your genus of interest and click on it (this is a work in progress and not all genera are clickable). In the introduction for that page, you will see a link to download my FASTA files of DNA sequences for that genus.

I’m trying my best to gather reliable sequences and label them properly, but many of my IDs are still theories, although backed up by research.

**2c. Finding reliable sequences of a species on the internet**

**Internet sequences are not reliable**

GenBank is the largest repository of sequences on the internet, holding sequences of every region of DNA of every living thing (plus viruses). UNITE is another large database of only the ITS region of DNA (the one I’m most interested in) of eukaryotes only (animals, plants, fungus and the like that have cells with nuclei). You might assume that since professional mycologists were the ones who put most of the sequences up there, that you can believe the ID of a mushroom sequence, but, no offense to anybody, mushroom identification is hard, even with DNA, and it’s often just not possible to be confident in an identification. Some studies have shown that the identity of a mushroom on GenBank is wrong 30% of the time. My own informal experience tells me that for some lesser known and large genera, GenBank is wrong far more than 50% of the time. UNITE is curated more closely and tries to only allow sequences with more confident IDs, but my own experience is that it is not that much more accurate. In other words, **you should assume that the identity of the mushroom is wrong**.

**Type sequences are the most reliable**

A type collection is an official collection of a species of mushroom. The **holotype** is the original, first designated collection of a species, guaranteed by definition to be that species, although if a few mushrooms growing together were picked at the same time and designated as the official type collection, and one of them turns out to be different than the others, that complicates things. The holotype is definitely the real thing: if there was a mixup and a million people and a hundred guidebooks all say a certain mushroom is one thing, and we later find out the holotype is of something different, all the guidebooks will have to be rewritten (unless a special exception is voted on) - the holotype can never be wrong. Unfortunately, we only started keeping dried specimens in herbaria relatively recently, so many species don’t have a holotype. An **epitype** is a specific collection designated later to be the official one because there was a problem with the holotype - perhaps it couldn’t be sequenced because it was very old or it was just a picture somebody drew. A **neotype** is a specific collection designated later to be the official one because there never was a holotype. A sequence of any of these three is about as good as you can get to being a reliable sequence of a species. An **isotype** is a different collection, but one that the person who described the mushroom is very sure is the same thing, perhaps from another location far away. Because mushrooms that look very much alike but growing in a different part of the world are often turning out to be different species, an isotope is not 100% reliable to be the same mushroom, but an isotype sequence is still very valuable.

**but we don’t have very many**

Unfortunately, we only have type sequences for less than 1% of mushrooms, so you are not going to be able to find one very often.

**How to find a type sequence**

To make matters worse, type sequences aren’t always labeled as such in GenBank, so there’s no way (yet) to reliably find them even when they exist. Here are some things you can try:

* Ask GenBank for a type sequence of something by including “type material” in your search, for instance “Cortinarius violaceus type material”. That will find the neotype sequence in this case, but more often than not, even when there is a type sequence, this method will not find it.
* You can read all the papers and manuscripts you can find on a certain group of mushrooms and look in those papers for a mention of what the type sequence is of a species. I google “pdf Cortinarius violaceus phylogeny” and start reading the papers that come up to see if one mentions what the type sequence is.
* You can also look on IndexFungorum. Visit [www.indexfungorum.org](http://www.indexfungorum.org), click on “Search Index Fungorum” on the top right, type “Cortinarius violaceus” in the search box, click on “Cortinarius violaceus” on the top left hand side of the matches that come up, notice that it has an older name and click on “Basionym Agaricus violaceus”, and then you’ll see that the neotype is “Moser 74/208”. Go to GenBank and search for “Cortinarius violaceus Moser 74/208” and the type sequence will come up, NR\_173726. We got lucky that GenBank labeled it the exact same way as Index Fungorum. Usually this does not work.

The mycological community acknowledges that getting type sequences of all the species is probably our #1 taxonomic priority and is the bottleneck keeping identification using genetic methods from being practical, but we’re a long way from getting them.

In our example case of *Cortinarius violaceus*, both methods above succeeded in obtaining a reliable sequence of it, and we can stop and proceed directly to step 3. But we got lucky. Usually, you can’t find a type sequence and you have to research further.

Let’s say we couldn’t find a type sequence of the species we’re interested in.

**Where was the mushroom described from?**

Arguably, the most important thing to know about a mushroom species is where it was described from, and was it from Europe, eastern North America, western North America or elsewhere. A mushroom not described from western North America may not occur here in western North America (where I am writing from), even if you find mushrooms here that look really, really, really the same, even microscopically. Ideally, you’ll want to know if a mushroom was described from your local ecological region.

First, go to [www.indexfungorum.org](http://www.indexfungorum.org) and click on “Search Index Fungorum” in the top right. Type “*Cortinarius violaceus*” into the search box. Find the result that best matches your species (maybe you are looking for a certain variety or subspecies, but maybe you aren’t) and click on the left hand name of the best match to bring up a page about that mushroom.

Graphical user interface, text, application, email

Description automatically generated

After the name of the mushroom, *Cortinarius violaceus*, you’ll see the names of people, perhaps in parentheses, then not in parentheses. Then you’ll see the name of a publication and a year. The person in parentheses, L. for Linnaeus, first described the mushroom with a different name. The different name he gave it is listed under “Basionym: *Agaricus violaceus* L. 1753”. So in 1753 Linnaeus described this mushroom. Then Gray renamed it to Cortinarius violaceus in 1821 in a publication called “Natural Arrangement of British Plants”. Linnaeus was Swedish and Gray was British, so we conclude this is a European mushroom, not an American mushroom.

To make sure, click on the basionym “*Agaricus violaceus*”. (That basionym may have another basionym that you may then have to click on, and so on, until you get to the oldest basionym). On that page you will find “Locality: Europe”. We have our answer. For Canada and the U.S., the Locality will often give a province or state, which is great.

Tips and tricks (if the locality field is not filled in):

* Did L. (Linnaeus) Fr. (Fries) or Pers. (Persoon) describe the mushroom? They mostly worked in Europe in the late 1700s and early 1800s so most of those mushrooms are European mushrooms.
* Was the date of the oldest basionym a date before 1890? Very few mushrooms outside of Europe were described before then so you can often assume Europe.
* Was your mushroom described by Peck? He mostly described mushrooms from eastern North America, usually NY state.
* How about Smith or Murrill? They could be anywhere in North America, east or west, so if it doesn’t say, you’ll want to read the publication to find out. For a very small percentage of mushrooms, mostly old ones, you can click on the “page image” link to read the publication, but due to copyright issues and logistics, that seldom ever works. Let’s hope that feature is prioritized in the future.

**Locally described mushrooms**

If that mushroom was described from your ecological and geographical area, in my case the PNW, then that mushroom does occur there by definition. It might not have been given the best name if genetically it does not belong to the genus it was assigned to. Whenever that happens, we should get the name changed to something more appropriate. If *Cortinarius violaceus* was described from the PNW, and we compared the dozen or so local sequences of it that we found, and they all appeared to represent the same genetic species, we could conclude that we now know the ITS sequence of *Cortinarius violaceus*. If there was enough variation in the dozen sequences that it wasn’t clear they were all the same species, we would want to investigate the possibility that there is more than one lookalike species that we don’t know about.

But the truth is, *Cortinarius violaceus* was described in 1753 (as *Agaricus violaceus*) and first collected in Europe. Thus, even though every guidebook here says it is found in the PNW, we have to verify that by comparing DNA sequences of European collections with our local collections to see if we really do have that species here or if ours is a cryptic sister species that just looks an awful lot like it (this happens all the time).

**Using UNITE**

**Question 1: Whenever somebody thought they had a certain species, how many different genetic species were actually collected?**

Let’s use UNITE <https://unite.ut.ee/search.php> to try and find reliable sequences of our species. It has a graphical map function that GenBank and BLAST don’t have.

* Type “*Cortinarius violaceus*” into the “Taxon name” box, but don’t press ENTER. The name of the species you are typing will show up in a pop-up box underneath where you are typing and you have to click on the name that pops up. If you just type the name and press ENTER, it won’t work.
* Next change the “Threshold” to 0.5%, and check the “Include sub-taxa” checkbox. Now click on “Go”.

Table

Description automatically generated

These results are telling us that when people sequenced something they thought was *Cortinarius violaceus*, all those sequences fell into 8 different groups, where within one group all of the sequences were within 99.5% the same (differing by at most 0.5% from each other). Each of those 8 groups is called a “Species Hypothesis”. If you get no results, perhaps UNITE knows the mushroom by a different synonym, so try all the synonyms. Or maybe nobody using UNITE has sequenced that species yet. You can try GenBank, discussed in the next section.

In other words, assuming that sequences within 99.5% of each other in ITS are the same species and that sequences that are less than 99.5% similar are different species and that all the sequences were clean and done properly (none of that is true, but we have to start somewhere) our results are telling us that **people found 8 different species of mushroom and thought that they were all *Cortinarius violaceus***. At most one of them can be right. So which one of the 8 species is the real thing? And will we need new names for the other 7? Perhaps nobody ever found and sequenced the real thing yet, so all of the choices are wrong. That’s always a possibility.

In this case, 63 of the 74 sequences agree, so #1 could be the real thing. Click on it. UNITE will bring up a map showing where those mushrooms were collected (when known). Don’t you wish GenBank could do that? Now you know why UNITE was created.

Graphical user interface, website

Description automatically generated

This species is found in Europe, where this mushroom was described, so it might be the real thing. It’s vital that this species be found near the type area in order to be a serious contender for being the real thing. That’s why we went to all that trouble up above to figure out where it was described.

We’re lucky and we can prove that this is the real thing, because we can see that the type sequence we found above, KM253741 (which I highlighted in blue) is found in the list of 63 sequences in that hypothesis. Do you notice that nowhere does it say it’s the type sequence? Wouldn’t it be really nice if it did?

If we didn’t already have the *Cortinarius violaceus* type sequence in our FASTA file, we would now get a reliable sequence from here. In the top left it says a good reference sequence is AY669579, so you could click on AY669579 in the left column and you will see the sequence. If there is no reference sequence, pick a good one. A good one will be one that is long, from the type area, and close to all the others (not an outlier). On the bottom right, if all of the columns are shaded green, including the left-most <.5 column, it is not an outlying sequence. If you can see letters visible under “Alignment” it might be one of the longer sequences, and you want the sampling area to be in Europe, where the type was collected. After you’ve chosen your own reference sequence, click on it to be taken to a page where you can copy the sequence letters. Copy them into your FASTA file for *Cortinarius section Cortinarius* and add a header such as this:

>Cortinarius violaceus UNITE EU 63/79 AY669579

ACCTAGTGCTA…etc.

In this case, I put the GenBank accession number, the name of the species and an indicator that it is an EU sequence from UNITE where 63 out of 79 sequences agreed (to justify why I think this is a real sequence of that species). For a complete explanation of how I label my sequences, read the appendix below “Sequence label syntax”.

Notice that this species is also found in North America. The circles mean “somewhere in Canada” and “somewhere in the US” (because of Alaska and Hawaii the circle that tries to show you the centre of the US is not actually drawn in the centre of the continental US). The dropped pins represent actual locations. The dropped pins show that this species is found in eastern North America, western North America and Europe. Often, those three areas have three distinct species that are lookalikes, but in this case, this same species is found in all those areas.

Let’s go back and check out the 2nd species hypothesis, with 4 sequences. They are from eastern North America. Hmm… back east they may have a second lookalike species. Somebody might want to investigate that.

Let’s go back and check out the 3rd species hypothesis, with 2 sequences. Those are from Norway. That’s close to Sweden, the type area. It’s possible these are the real thing, but clearly the species with 63 collections is more common than this species with 2 collections, so the odds are the first *Cortinarius violaceus* that was ever found back in 1753 was the other species.

Let’s go back and change the “Threshold” to 1.0%. Now look:

Table

Description automatically generated

Now hypothesis #1 has 72 species in it. All 4 sequences of old hypothesis #2 and both sequences of old hypothesis #3 were only barely >0.5% different, and they are all now included in a new hypothesis that includes everything within 1%. Perhaps those different eastern North American and Norwegian sequences were not that different after all.

**More than one possibility for the correct species hypothesis?**

We got lucky. We now think we know what a real sequence of *Cortinarius violaceus* is. But what if there were 2 species hypotheses, each with 50 sequences, both with many collections from the EU, differing by, say, 4% from each other with no clear winner? Then we can’t tell which is real, and we have to put both possibilities in our FASTA file. I will pick a reference sequence from both hypotheses and label one of them “>Cortinarius violaceus UNITE EU#1” and the other “>Cortinarius violaceus UNITE EU#2” and in my trees I’ll see the two possibilities of the real thing. This creates a mystery – which is correct and what is the other one?

**UNITE accession numbers**

If the accession number starts with UDB (UNITE database) then you can’t find it in GenBank, only in UNITE. You can type it in the “Accession Number” box on the UNITE search page, but you can’t type it into the GenBank search box, it won’t work. And you can’t use it in a BLAST window either.

**Question 2: Whenever somebody actually had a certain species, what were all the different mushrooms they incorrectly thought they had instead?**

In question 1 we talked about misidentification that happens when a bunch of people think they are collecting the same thing but not all of their DNA sequences match and we discover that more than one species has been collected under that name. Now, we think about it the other way around. We will now talk about misidentification that happens when the same actual species was collected by a bunch of people, but they all didn’t agree on the name they gave it. Let’s go back and look at that map screen again:

Graphical user interface, website

Description automatically generated

This is the list of all 63 sequences that match *Cortinarius violaceus* within 0.5%. Notice that towards the bottom of the image, one person in Germany thought they had collected *Cortinarius hercynicus*, a very similar species, instead. If there is more than one possible species hypothesis in UNITE, you should investigate all of the different names given to collections in all the possible species hypothesis as possible identities for the “other” species hypotheses that turn out to not be the real one.

After using UNITE as explained above, you now have a list of all of the mushroom species concepts (assuming 0.5% similarity means a species boundary) that:

* somebody thought was your mushroom
* collected in the right part of the world
* you also know what other names it was called when it was found, so you can see what species it might be confused with.

Now you can make an educated guess as to which sequences are identified correctly (reliable) and which aren’t. If using UNITE doesn’t provide a clear answer, you can also search through GenBank for reliable sequences, although the process isn’t as friendly.

**Find sequences in GenBank**

GenBank searching is explained above. Instead of using GenBank to search for local sequences, you can use it to search for sequences from the type area to get some ideas of what might be a reliable sequence, but you’ll have to look at the description of each sequence to try and determine its location as there is no mapping function. Remember to search for all possible synonyms of a species.

Collect every sequence of that mushroom you can find, and then make a quick BLAST tree of them all to see if all the mushrooms from the type area have the same sequence or not. If so, you might be able to a long, healthy, reliable reference sequence you can use.

**3. How to compare your sequence to others**

Now you have your sequence of interest, and reliable sequences of some of the possible species it might be. It’s time to compare two sequences to determine if they represent the same species or not. If you don’t really know what genus you have, and all your files are split into a separate FASTA file for each genus, you may have to make a big FASTA file combining a bunch of different genus FASTA files and BLAST against that, or try different possible genera, one after another.

**BLAST against my or your own reliable sequences**

My reference sequences are found at [www.alpental.com/psms/ddd](http://www.alpental.com/psms/ddd). Search for the likely genus of your mushroom and click on that page (only some pages have been written so far, this is a work in progress). In the Introduction section at the top, you will find a “link to all my FASTA data” for those mushrooms.

Visit BLAST at <https://blast.ncbi.nlm.nih.gov/Blast.cgi> and click on “Nucleotide BLAST”. On the BLAST page, click the checkbox that says “compare two or more sequences”. Now there will be two big boxes you can paste sequence data into. In the top box, put the single sequence that you are trying to identify. In the bottom box, put the list of all the reliable sequences you want to compare it to. If there are more than 100 of them, you will have to go open the “Algorithm Parameters” at the bottom increase the “Maximum target sequences” to 1000 (or however many you might have). Some of my files might have up to 1000 sequences, and if you don’t change this number, BLAST will only look at the first 100 sequences in a file, and ignore the rest, and your BLAST won’t work.

**What % similarity indicates the same species?**

A sequence that is less than 97% similar in ITS is likely a different species (< 1% of the time will it be the same). From my own experience, to get a >50% chance that two sequences are the same species, they have to be within 99.5% of each other in ITS. The percentage identical you are given is an APPROXIMATE only. You have to investigate the comparison by hand to determine the real % difference before you can reach any conclusions. I’ll discuss that shortly.

If my sequence matches a reliable sequence truly better than 99.5% after examining the differences by hand, I will usually say “they are likely the same species”. The entire ITS region is about 600-700 characters long, which means that 0.5% difference is a difference of 3.5 characters. In other words, if the sequences differ in 3 or fewer places, I lean towards thinking they’re the same species. If they differ in 4 or more places, I lean towards thinking they’re different.

The 99.5% figure is an approximation and does depend on what genus it is… for some groups of mushrooms, two sequences may typically have to be >99.75% similar to be the same species. For other groups, it seems 98% is close enough, so it’s good to talk to an expert in the group about it. Also, you could look at a tree of that group of mushrooms and see how far apart neighboring species typically are. My web page also has trees to look at for each group (or you can make your own trees using my tree tutorial “How to make genetic trees”).

Note that my figures of 99.5% and so on are only valid for ITS sequences, a “junk” region. If you are comparing LSU, SSU RPB2, etc. or any region that encodes for proteins, the sequences have to be even closer than that to represent the same species (practically identical). Such regions don’t have as many differences between genera as ITS does, so they are better to place a species to its proper genus or family. ITS is best for determining if two collections are the same species or not, which after all, is the goal of this tutorial.

You may need to find and sequence dozens of examples to see if they clade together in a few clumps or are spread out by a few % points with no clumping to give you an additional data point of if they might be different species (the former case) or one plastic species with more genetic variation than usual (the latter case).

\*\*\* new INTERPRETING TREES tutorial needed \*\*\*

**DNA is only one tool of many**

Of course we want to determine something definitive, but it’s important to know that DNA is only one tool for determining species, albeit an important one, but you can’t reach a conclusion without using other tools as well. You should evaluate each mushroom in at least 3 categories:

1. Morphology, both macro and micro
2. Ecology
3. DNA

If your mushroom matches the description of a species in most significant ways, as well as matching the microscopic descriptions in most ways, and it has the same ecology (grows near or on the same kind of tree, or at least the same group of trees (eg. conifer vs. deciduous) then that adds weight to them being the same species, even if the genetic variation is > 0.5%. If they don’t, then it’s possible that you could have a 100% match in ITS and still have a different species (that happens more often than you might think).

The final call determining if two sequences represent the same species cannot ever be made without knowing the morphology and ecology of both collections.

**Comparing two sequences by hand**

After performing a BLAST (see section 2a), click on one of the results of the comparison to see the details. Now do a Ctrl-F or Command-F to search the page for “| “ (vertical pipe followed by a space). That will highlight the differences, except for differences at the beginning of a column, so make sure to scan down the first column to look for differences that aren’t highlighted. Here is an example:

Text

Description automatically generated

The top section is ITS1, a “junk” region that is subject to random mutations. The middle section, which is usually the same between species, is called “5.8s” and you can expect little to no difference between them as it is not a junk region, and any changes that happen there affect the ability of the organism to reproduce and therefore can’t get passed on. The bottom section is ITS2, also a “junk” region subject to random mutation.

Table

Description automatically generated

1. There are 4 kinds of nucleotides, A C G and T. If you have a different letter in one sequence than in the other, you have a base pair difference, or bp. You can see several of them above.
2. You might see a dash “-“ in one sequence and a letter in the other. This is an insertion or deletion (indel) that means one sequence is shorter than the other and is missing something at that position. Not every indel is real.
   1. It’s sometimes hard for a computer to tell how many of the same letter are in a row. In the top row you see 4 Ts in the query sequence and 5 Ts in the subject sequence, and in the bottom row you see 5 Gs in the query sequence and 4 Gs in the subject sequence. Perhaps the computer just had trouble telling if there were supposed to be 4 or 5 of them. You can look at the original chromatogram graphs for both, if you have them, and find out.
   2. When a single letter is repeated a lot (or a pair or more of letters repeated a lot) and that DNA is being copied, it is very susceptible to errors (leaving out a copy or adding an extra copy). A change from 4 to 5 repeats of a string could happen so easily, that’s another reason that it should not be counted as a real difference between the two sequences. So, if one sequence has “ATATATAT----" (4 of them) and the other has “ATATATATATAT” (6 of them) that may not be a significant difference.
3. Your sequence may have what is called “**ambiguous locations**” or a “**polymorphism**”. This means that you might see an unusual letter of the alphabet in a sequence, like K, R, S, W, M or Y. Google “nucleotide wiki” and scroll near the bottom of the wiki page to see the definitions of the other letters. For instance, in the bottom row above the last difference is an R on top and an A on the bottom. Mushrooms are diploid creatures like humans, which have 2 complete DNA strands inside them, and where they differ our features are determined by considering both of them. For instance, An R means that one copy has an A and the other copy has a T in it. It seems both versions exist in the population, so the fact that one collection had the A in both places, and the other collection had an A in one copy and a T in the other copy, might not matter. I assume that any difference where at least one of the two sequences has an unusual letter (R T …..) is not a real difference.
4. If you see an N, that means that the computer was unable to figure out the DNA nucleotide at that spot. There is no reason to assume it would have differed from the other sequence, so that is probably not a real difference.
5. Where there are 2 missing nucleotides in a row (more than one “-“ in a row), an entire section might have dropped out or been inserted in one event, so you can count the whole “chunk” as one difference. In the second row, I count three missing letters as one difference, not 3.

In the above diagram, what looks like 10 differences, I only count as 4 bp differences and 1 indel (5 differences). These two species are more closely related than it first appeared. If the computer said they were 99.0% identical, they might actually be 99.5% identical.

A picture containing letter

Description automatically generated

1. Mutations happen randomly, so if you see there are a number of differences fairly close together instead of randomly scattered throughout the sequence, that doesn’t sound real. That could mean something went wrong (think a smudge on the glass plate) and the computer got the wrong result in that whole area, and you might be able to ignore all those differences in the area where many changes were detected, and only consider the scattered differences you see elsewhere as real. In the above example, the top area with many differences is suspicious, but the few scattered differences below that seem legit. I would try and find the original chromatogram graph for them both and check how clean the graph is for that top line to see if indeed that section was dirty and read incorrectly for one of them.

A screenshot of a computer

Description automatically generated with low confidence

1. It’s hard to know how far apart consecutive nucleotides are at the very beginning or end of the graph, since the x-axis isn’t quite linear at the ends, and it’s possible for the computer to be unsure if there is 1 or 2 copies of the same letter in a row. This happens around 7 positions or so from one of the ends and always where one copy thinks there is an extra copy of a letter and one doesn’t. In the above example, this is not a real difference. The incorrect copy will be the one close to one end of the read. In this case the subject sequence is probably wrong, since the query sequence started earlier and by the time it got to position 38 the graph would be quite linear and unlikely to be mis-read.

Text, letter

Description automatically generated

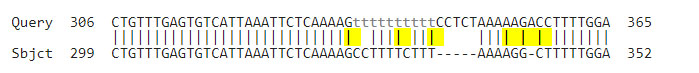
1. As also explained in my tree tutorial in section 1, if at least one of your two sequences doesn’t start or end at 1, that’s because there were no similarities in the first chunk of the sequence, so your sequences are more different than you think. If the first sequence starts at 53 but the second sequence starts at 3, then the first two locations of the second sequence were different, and you have to add two differences to the count of differences you are seeing to get the true number of differences. If the lowest of the two numbers is 10 or higher, for instance, the sequences are so different you should not even compare them, and just assume that they are different species. In the above diagram, the first and last numbers of the query and subject are 486, 251, 656 and 421. The lowest of those numbers is 251. That means that up to 250 positions are not shown because they are mostly different. So there may be an extra 250 differences between these sequences besides the 11 differences that the computer is showing (in the middle sections where there were some similarities).

A picture containing text

Description automatically generated A picture containing text

Description automatically generated

1. Here is an example of figuring out that a sequence is corrupt and can’t be trusted. I am comparing a known good query sequence to two similar subject mushrooms, to see if they are the same or a closely related sister species. **On the left**, we see that the subject mushroom is not the same species, but has a number of differences in ITS1 and ITS2. The 5.8s region where there should not be any significant differences runs from about query location 220 to 350. But now look **on the right**. Not only are there some differences because the subject sequence has some N’s in it, where the sequence could not be read, but we have to ask ourselves if the problems reading the sequence were so bad that it did not produce the correct letter in a lot more places, but thought it did. Can we trust the parts of the sequence that are not Ns when so many Ns are present? In this case, no. The 5.8s region from query location 220 to 350 is full of differences, and if this mushroom is even in the same family or order as the other two, which we know it is, there are really no differences in that region. So DO NOT TRUST subject 2. Don’t waste your time looking for a matching sequence somewhere that does not exist - throw the sequence away.



1. Sometimes a difference can be interpreted in more than one way. After the long indel where a chunk of 5 is missing on the bottom, it looks like there are 2 bp differences and another indel in between them. But that could also be explained by one indel (5 A’s on top and 4 on the bottom, which as explained above is probably not a real difference), then one indel (1 G on top and 2 G’s on the bottom), another indel (one A on top and no A’s on the bottom), another indel (2 C’s on top and 1 C on the bottom), and finally one last indel (4 T’s on top and 5 T’s on the bottom, also probably not a real difference). This interpretation is a difference of 3 indels, versus what the computer is showing (2 bp differences and one indel). That’s still 3 total differences, but as bp differences are probably more significant than indels, and some of those indels may not end up being a real difference once you look at the chromatogram, this tells me that the differences in that area may not be as significant as the computer is showing.

**Example: Is Cortinarius violaceus found in the PNW?**

It’s time to find out. In step 1 we found 2 sequences of Cortinarius violaceus from Oregon, MH266404 and MK949107. In step 2 we found the type sequence, NR\_173726. Now let’s compare them using BLAST by putting the type sequence in the top box and the two Oregon sequences in the bottom box.

Text

Description automatically generatedText

Description automatically generated

On the left, we have 3 bp differences that look real, 2 in ITS1 and 1 in ITS2. One of the ends of the sequences starts at 1 (the subject) so there are likely no additional differences not shown. 3 out of 656 is greater than my usual threshold of 99.5% similarity to be considered the same species.

On the right, we have 6 bp differences, but one of them is an N so that isn’t real. With 5 differences (2 in ITS1 and 3 in ITS2), that is still < 99.5% identical, so I start to wonder if that isn’t actually *Cortinarius violaceus*. If no PNW collections were >99.5% similar, I would doubt it even more, but we see with the one on the left that PNW material is capable of being within 99.5%. The one on the left is within 0.5% of the type sequence, and the one of the right is within 0.5% of the one on the left. If you look at a tree with many sequences from all over the world in it, the sequences are scattered about the tree all within 1% of each other or so, there are no “clumps” of sequences like you would expect to find if there was more than one species (with empty regions between them with no sequences).

So what does it mean? Others have studied material from all over the world and there are no ecological nor macroscopic nor microscopic differences of note between our collections and EU collections, so we officially, for now, have Cortinarius violaceus in the PNW and not an undescribed sister species. If those studies had not already been done, this might indicate that we should take a careful look at their ecology, etc. to make sure they are both the same species.

**Check for existing synonyms of a new species before naming it yourself**

Consider the example of *Hygrocybe miniata*. Local sequences turned out to be 4-5% different than UK type area sequences. So our species is probably not that. But does it need a new name or does it already have one? Search for *Hygrocybe miniata* in Index Fungorum (IF). Then click where it says “also see Species Fungorum: Hygrocybe miniata” to see a list of names that Species Fungorum (SF) considers synonymous. Search for each synonym in Index Fungorum to see if any of them were described from your region, in my case the PNW. There are none. But others have suggested there are more synonyms. MycoMatch ([www.mycomatch.com](http://www.mycomatch.com)) is a free PC program (of which I am a co-author) containing all that is known about the mushrooms in the PNW. That program says that Hesler and Smith’s 1963 book “*Hygrophorus of North America*” lists *Hygrophorus constans* as a synonym, which IF says has an older basionym of *Hygrocybe constans* which IF says was described from Oregon. Thanks to MycoMatch, I didn’t have to go reading every book on the Hygrophoraceae to figure this out.

Murrill described *Hygrocybe constans* from Oregon in 1912 as being different than *Hygrocybe miniata* in the fact that the red colour didn’t seem to fade in age like it did in the EU species. Hesler and Smith figured out that the amount of colour fading is highly variable both here and in the EU, so Murrill’s name was synonymized with *H. miniata*. But now we have evidence that our local species varies by enough in ITS DNA (4-5%) as to probably be a distinct species. I have dozens of local collections of it. The sequences seem to be 1% or more apart from each other within the PNW, but there are sequences with ambiguous characters that show that not all the differences may be meaningful. If you make a list of all the places that can contain an ambiguous character, the number of places that are still different is small enough that the sequences may all be within 99.5% of each other.

Chances are, Murrill picked the same thing as the dozens of collections that have been sequenced, as no other mushroom has been found here that looks like it but has a different sequence profile. Murrill was right that it was a new species, just perhaps for the wrong reason. Now we have to look closer at it to figure out how it actually does differ from the EU *H. miniata*. (Maybe it does hold its colour longer, who knows). But the biggest differences are probably still unknown. *Hygrocybe constans* is probably already the correct valid name for our local species, long mistakenly called *H. miniata*. Giving it a new name without doing this research to find out that it already had one would have been a mistake. And I figured all of this out from a laptop with an internet connection, and you can too.

**Here are some guidelines I go by:**

* If two type sequences (or very reliable non-type sequences of two different species) have the same (or very similar) ITS DNA, look at their descriptions. Were they separated on the basis of the thickness of one kind of cell wall? Maybe they don’t deserve to be separate species given this new information that ITS DNA does not differ. Were they separated on the basis on one growing with conifers and the other with deciduous trees? Maybe they do deserve to be separate species.
* Do two sequences of supposedly the same species differ by > 3%? Maybe there is more than one “cryptic” species here. Get several collections of each and look for differences between them.
* Are the results ambiguous (between 0.5% and 3% different) between type area sequences and your local sequence? You can make an initial educated guess on whether they are the same or different this way - was there ever a local name given to this mushroom (like *Hygrocybe constans* for *Hygrocybe miniata*) where some people once thought our local mushrooms was different (Murrill) and some other people thought they were the same (Hesler and Smith)? And do the differences found by the first person seem real? (Ecological and not the thickness of one particular kind of cell wall, for example). If so, my initial guess will be that our species might be different, since somebody once found differences. If not, and the only known difference is the DNA difference you found, that’s not enough, so my initial guess will be that our species is the same.

**Appendix - Sequence label syntax**

Here is how to interpret the labels I have given to my sequences:

EU = Europe. NA = unspecified North America. ENA = eastern North America. WNA = western NA. WA, BC, PQ = two letter abbreviation code for a state or province.

Hebeloma mesophaeum TYPE EU AF621123 - this is a type sequence from the EU, which is the most reliable kind of sequence you can have. Sometimes I specify HOLOTYPE or EPITYPE or NEOTYPE (the best kind) or ISOTYPE (the next best kind). I always use capitals for the word TYPE when it is the type sequence itself. Two letters followed by 6 numbers means that is the GenBank identifier, so you can use that to find more information about a sequence. If the sequence identifier starts with UDB, then you have to find more information in the UNITE database. (This is only an example, that is not actually the Hebeloma mesophaeum type sequence).

Hebeloma mesophaeum Beker EU (type EU) - this is a sequence provided by a publication by Henry Beker, an expert in Hebeloma, from the EU, and the lower case “type” means that the type collection is from the EU, so the sequence is from the correct general area and vouched for by an expert. Many sequences in many papers by many authors are incorrect (and they don’t make any claim that they are necessarily correct), so I will generally only accept that one is reliable if the study actually confirmed the accuracy of the sequence identity. Or, I might show different possibilities by different authors.

Hebeloma mesophaeum WA? – A question mark indicates my best guess that can’t be confirmed. In this case, I think the sequence is of a collection from WA.

Hebeloma ‘mesophaeum’ WA – Single quotes around a name means it’s not actually that, although that’s probably what most people would think it was, to look at it. Single quotes around the genus means it needs to be moved because it doesn’t actually belong in that genus.

Hebeloma mesophaeum WA (ITS2) – this is not a full ITS sequence, it is ITS2 only (I also indicate ITS1 for ITS1 only).

Hebeloma mesophaeum WA AY619234 (as dunense) – This turns out to be H. mesophaeum, but if you look at the GenBank entry for it, AY619234, you’ll see they called it H. dunense because that’s what they thought it was until my or some other recent study determined the real identity.

Hebeloma mesophaeum UNITE NA #1 (type EU) - this is one of several species hypotheses found in UNITE, but the sequence is from North America, and the species was described in the EU, so it is not to be trusted.

Hebeloma mesophaeum UNITE EU 15/17 – 15 of the 17 UNITE sequences in all the species hypotheses of that name were in this particular species hypothesis. There were 2 sequences that didn’t agree, where the collection might have been identified as this, but sequenced to be something >0.5% different.

Usually when I label a sequence with a name, I believe the sequence to be correctly identified, unless I qualify it by something like “UNITE #1” (meaning it’s the first UNITE possibility) or “GenBank ENA #1” (meaning it’s one possibility from eastern North America found in GenBank) or by giving the name of the person who identified it.

Hebeloma mesophaeum WA (4/2 UNITE #1) - this WA state sequence is 4 bp and 2 indels different than the sequence I labeled UNITE #1. Take that with a grain of salt, usually that means out of 700 base pairs, but these sequences may be longer or shorter, so that doesn’t really tell you the % difference.

Hebeloma mesophaeum WA (4/2/3 UNITE #1) - this WA state sequence is 4 bp and 2 indels different than the sequence, plus, if I give a third number, that means that there are 3 ambiguous locations where an A was against a G, for instance. As discussed above, these are not necessarily real differences, but they can explain why a sequence appears to be far apart in a tree from other sequences of the same species. When there are ambiguous characters, the tree drawing algorithm may show the two sequences far apart, whereas one interpretation is that those differences do not matter as much.

Hebeloma mesophaeum WA (4/2/2n UNITE #1) – this WA sequence is 4 bp and 2 indels different than my sequence labeled “UNITE #1” but there are also two ‘N’s in one of the sequences, meaning a base could not be read and can’t be compared. It’s unlikely that represents a real difference, especially if there are few other differences between them. I assume the Ns don’t represent a real difference, but it’s always possible that a difference could just happen to be in a spot that couldn’t be read.

Hebeloma mesophaeum WA (3-4/0-2 UNITE #1) – this differs by between 3-4 meaningful bp and between 0-2 meaningful indels. I was not sure if a certain difference was important or not, so I’m providing a range of possible meaningful differences.

If I use the word “dirty” or “+Ns” that means there are unidentifiable characters (Ns) or another reason to believe the sequence is dirty. It won’t be in its proper place in the tree because of this, and it won’t match well to other sequences of the same species.

Hebeloma cf mesophaeum WA - “cf” means “confer” which means “compare with”. In other words, this WA sequence looked like H. mesophaeum, but I can only confirm a superficial resemblance. It’s probably not that (or at least there are competing possibilities) or I would have used stronger language.

Hebeloma aff. mesophaeum WA (7/3) - “aff” means “affinity”, which means it is closely genetically related, but probably not that. I actually have DNA of what I believe is the real thing, and this is probably between 0.5% and 5% or so, too different to be the real thing, but close enough to be related. I might have something like (7/3) in the label saying exactly how different it is from the sequence that is probably the real thing, in this case 7 bp and 3 indels different.

Hebeloma mesophaeum UNITE #1 EUx2 (+sp x24)

Hebeloma mesophaeum UNITE #2 EUx2 - 26 total sequences in UNITE matched the first sequence, but only two of them were labeled as that species from the EU. The other 24 were not identified specifically as that species, or they were not from the EU. There is also a competing sequence found only twice in the EU. What does this mean? sp. #1 is more common, since its DNA was found 26 times vs. only twice, but that doesn’t mean it’s the real species. Two times each that somebody thought they found H. mesophaeum, it was one genetic species or the other, so that’s a tie. If one of them was labeled with that specific name much more than the other, you could make an assumption that was the real one. Given that #1 is more common by far, you could make an argument that we should accept the more common species as the real thing, but that’s only an assumption and a guess.

Hebeloma mesophaeum WA iNat43123425 - you can visit [www.inaturalist.org/observations/43123425](http://www.inaturalist.org/observations/43123425) to actually see the mushroom and read about it. Collected in WA. I appear to be confident that it is H. mesophaeum, so you will probably see matching sequences in the tree that are a type sequence or other reliable sequence to explain why I am confident about the ID.

Hebeloma mesophaeum OR MO12345 - you can visit [www.mushroomobserver.org/12345](http://www.mushroomobserver.org/12345) to actually see the mushroom and read about it. Collected in OR.

Hebeloma mesophaeum PQ MP123456 - you can visit [www.mycoportal.org/portal/collections/individual/index.php?occid=123456](http://www.mycoportal.org/portal/collections/individual/index.php?occid=123456) to actually see the mushroom and read about it. Collected in Quebec.

Pyrrhulomyces astragalinus EU (+ID+WA) – This sequence is from the EU, but the same sequence has been found in Idaho and Washington (I may or may not put those sequences in the file too).

Pyrrhulomyces astragalinus 1821 EU (=Gymnopilus laeticolor 1912 WA) – sequences from WA of G. laeticolor are the same as this sequence from the EU of P. astragalinus. It is possible the two species are synonyms. If so, the former was described almost 100 years earlier, and as the oldest name, would take precedence. Usually, somebody will study the ecology and morphology of the two species and confirm this.

Pyrrhulomyces astragalinus EU (same ITS as Gymnopilus laeticolor WA) – if I word it this way, a study has probably shown that the two species have different ecology or morphology, so even though the ITS DNA is the same, they are probably being considered different species that cannot be told apart by DNA sequencing of the ITS region.

Hebeloma mesophaeum OR (longer type match) – the type sequence was very short, and doesn’t tell us what the entire ITS region is for that mushroom, so I found a longer sequence, this one, that matches the TYPE in the overlapping range but provides the rest of the data for ITS about this species as well. You may want to compare your mushroom to this sequence instead of the type sequence for a better comparison. It is common for old DNA to not sequence properly, so you might only get part of ITS1 or part of ITS2 out of an old type collection.