Jun 11, 2025

De Novo Assembly of Sequences from Eurofins with Geneious

DOI
https://dx.doi.org/10.17504/protocols.io.eq2lyn5wqvx9/v1

Dakota Betz[1]

[1]ucsd

Rouse Lab

Sonja Huc
UNCW

DOI: https://dx.doi.org/10.17504/protocols.io.eq2lyn5wqvx9/v1

Protocol Citation: Dakota Betz 2025. De Novo Assembly of Sequences from Eurofins with Geneious. **protocols.io** https://dx.doi.org/10.17504/protocols.io.eq2lyn5wqvx9/v1

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

Created: June 24, 2022

Last Modified: June 11, 2025

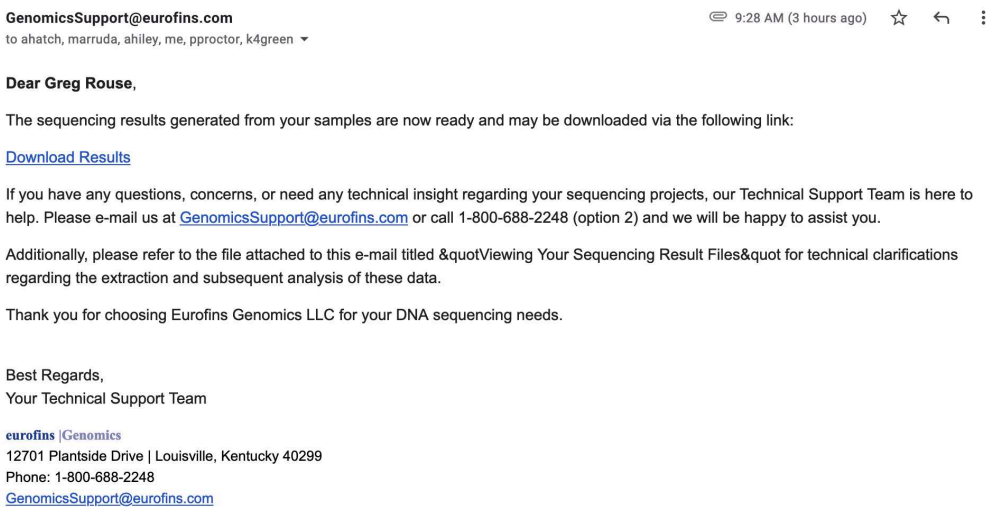
Protocol Integer ID: 65271

Keywords: eurofins with geneious, de novo assembly of sequence, de novo assembly, eurofin, geneious, sequence

Abstract

Our protocols are constantly evolving and old versions will be deleted.
The documents here are not intended to be cited in publications

1 Download the results from Eurofins. You should get an email like this:



GenomicsSupport@eurofins.com
to ahatch, marruda, ahiley, me, pproctor, k4green

9:28 AM (3 hours ago)

Dear Greg Rouse,

The sequencing results generated from your samples are now ready and may be downloaded via the following link:

Download Results

If you have any questions, concerns, or need any technical insight regarding your sequencing projects, our Technical Support Team is here to help. Please e-mail us at GenomicsSupport@eurofins.com or call 1-800-688-2248 (option 2) and we will be happy to assist you.

Additionally, please refer to the file attached to this e-mail titled &quotViewing Your Sequencing Result Files&quot for technical clarifications regarding the extraction and subsequent analysis of these data.

Thank you for choosing Eurofins Genomics LLC for your DNA sequencing needs.

Best Regards,
Your Technical Support Team

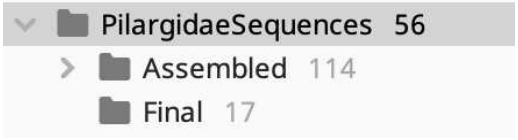
eurofins |Genomics
12701 Plantside Drive | Louisville, Kentucky 40299
Phone: 1-800-688-2248
GenomicsSupport@eurofins.com

Click on the underlined and blue highlighted "Download Results" in the second line of the email content. This will download a .zip file of the sequencing results.

Then, to keep your computer organized, move that file from your downloads into whichever folder you've designated for it, i.e. Documents > RouseLab > EurofinsRawReads > "DateOfResultsReceived".

Double click on the zip file to expand it.

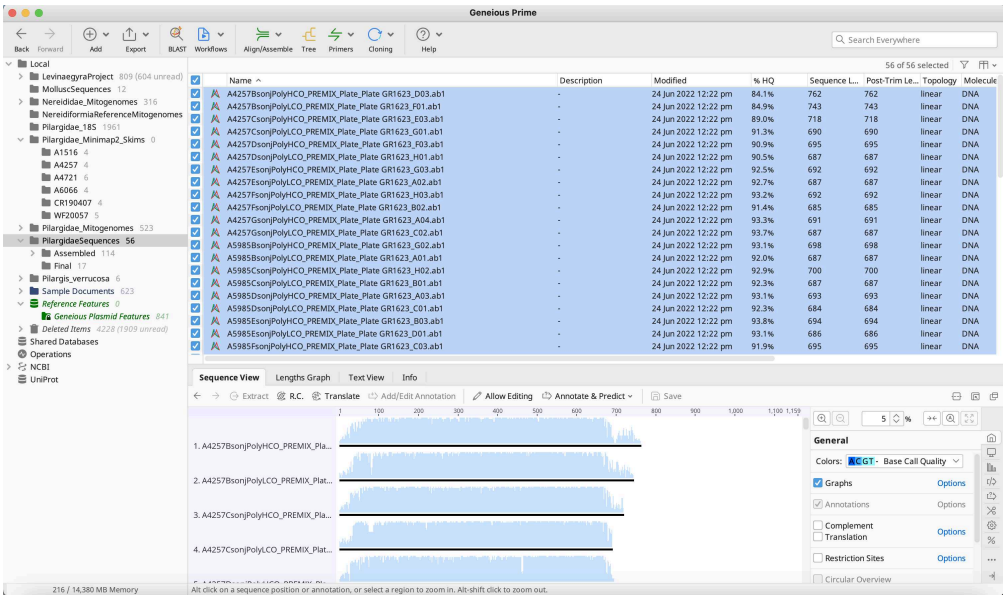
2 Open Geneious (either Prime or the shared license, or on a lab computer). Either navigate to or create a folder named "OrganismYouAreWorkingWithSequences" - mine, for instance, is "PilargidaeSequences". Within that folder, I also have an "Assembled" folder, which is where I put the Eurofins raw reads once I'm done with them, and a "Final" folder, where I put the sequences I get from the raw reads.



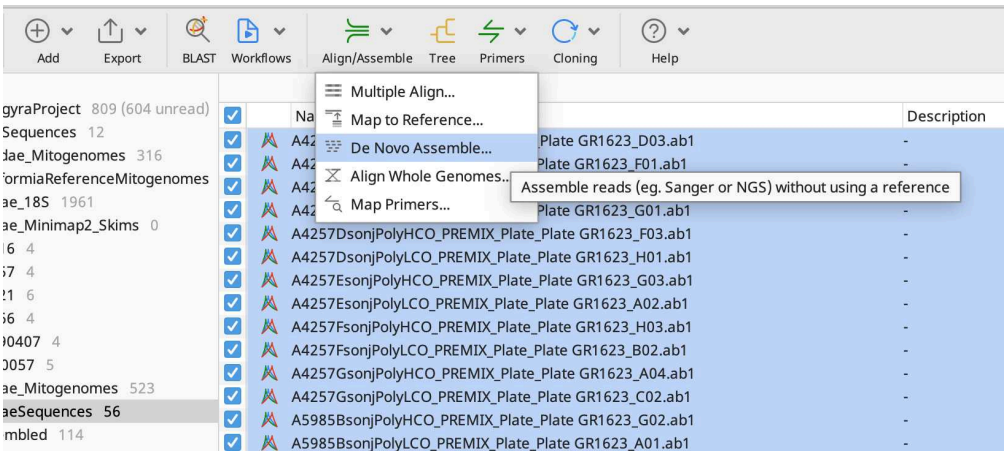
PilargidaeSequences 56
Assembled 114
Final 17

3 In the unzipped folder you downloaded from Eurofins, sort it by file type. You only want to import files with the ending ".ab1". If your computer allows it (mine doesn't like this, but it should work), you can also search for your four letter code (for instance mine is sonj, Marina's is mari, whatever your specific code is) within the folder to filter out the sequences that are yours.

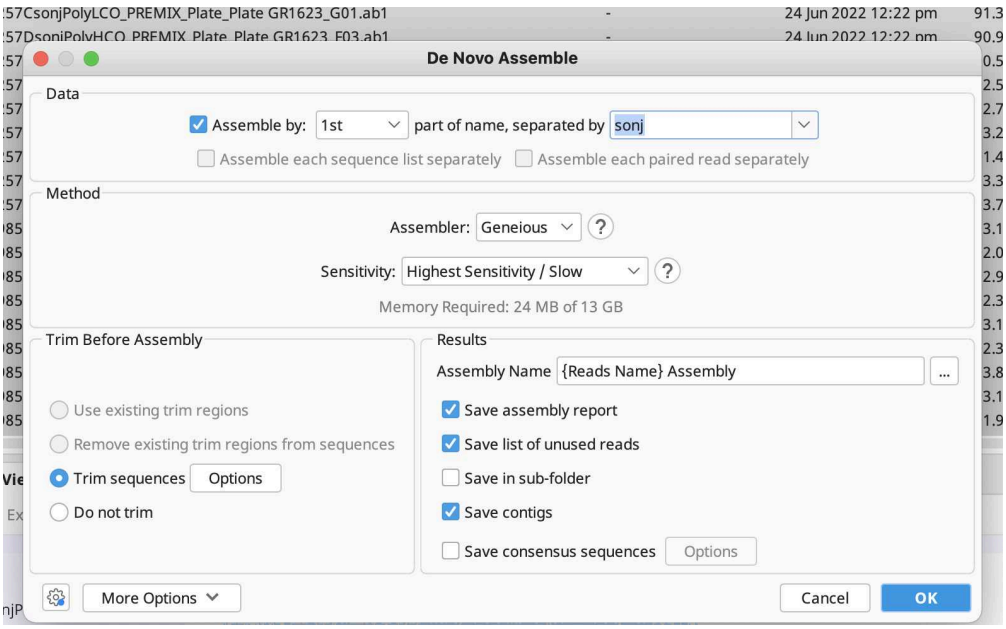
4 In Genious, navigate to the working folder you want to drag your sequences into. Then, in Finder, select the raw reads that are yours and end in .ab1, and drag them into Geneious. The app should look something like this:



5 If they aren't already, select all of the raw reads (you can do this easily by clicking the box just to the left of the "Name" field). Then, click on the "Align/Assemble" tool, and select "De Novo Assemble..."



6 Then, in the first part of the window that pops up, change the "part of name, separated by" to whatever YOUR four letter code is, and click "OK". The rest of the parameters should be default, but check to make sure yours looks like this, too.

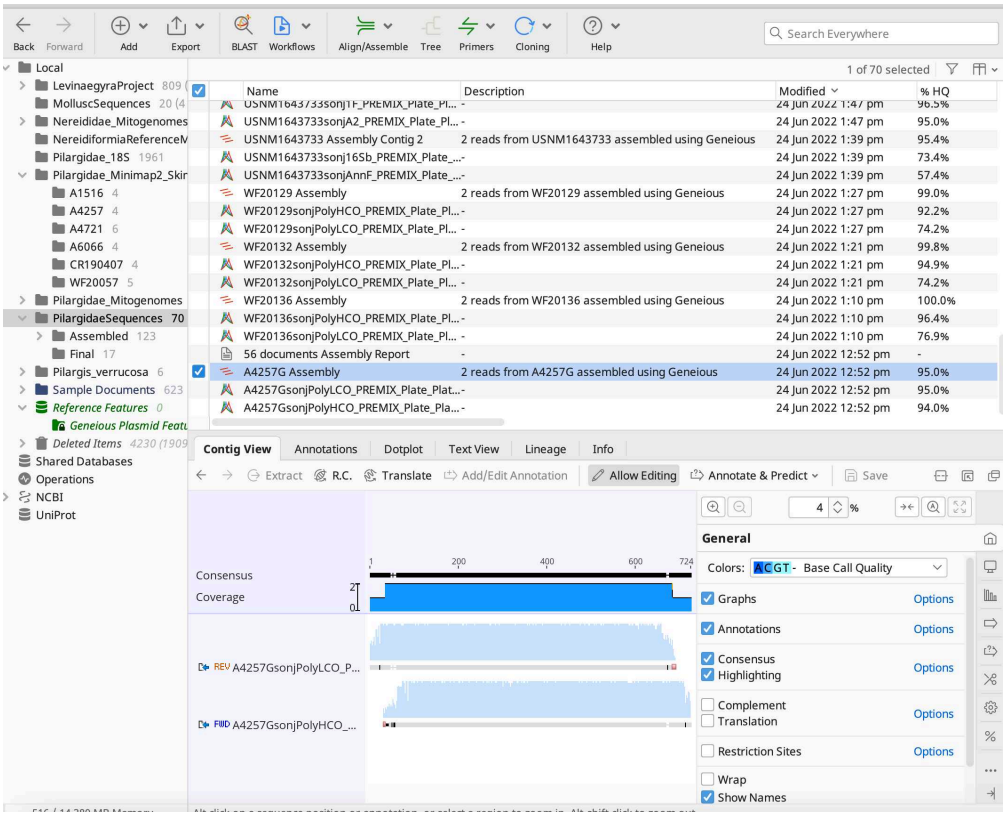


7 Create an excel file and create a table for your sequences. Mine is called "PilargidaeDeNovoAssemblyStats". Have column headers that are: "Sample ID", "File Name/Assembly", HQ% Before", "HQ% After", "Sequence Length Before", "Sequence Length After", "Ambiguities Before", "Ambiguities After", "Gaps Before", "Gaps After", and "Date Received".

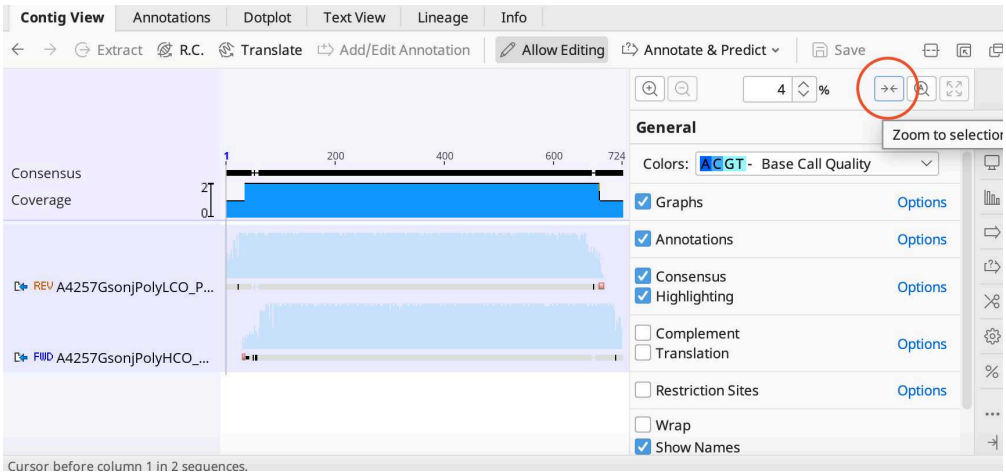
Mine looks like this I like to also add the stats for each raw read, not just the assembly, but the assembly by itself should be enough.

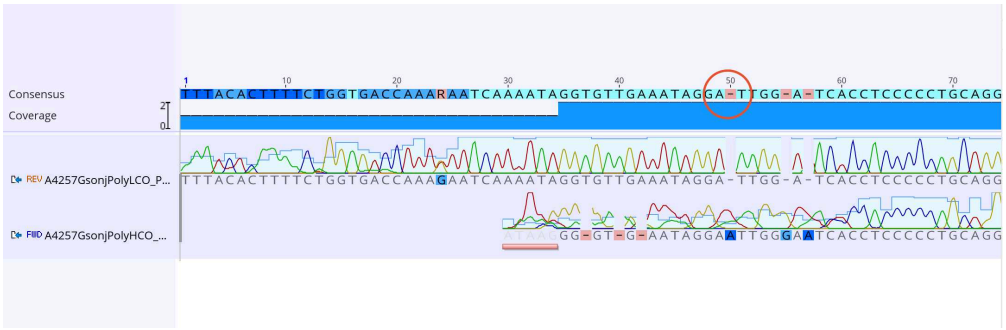
Sample ID	File Name/Assembly	HQ% Before	HQ% After	Seq Length Before	Seq Length After	Ambiguities Before	Ambiguities After	Gaps Before	Gaps After	Date Received
A2813	A2813 Assembly Contig arL/BrH	91.5%	99.3%	572	449	3	0	3	1	
	A2813sonjBrH_PREMIX_Plate_Plate 01_C03.ab1	49.8%		548		0		0		Aug 6 2021
	A2813sonjarL_PREMIX_Plate_Plate 01_A03.ab1	82.3%		552		1		0		Aug 6 2021
	A2813 Assembly Contig 18S	96.5%	99.0%	1,841	1,735	1	0	1	1	
	A2813sonj1F_PREMIX_Plate_Plate 01_G04.ab1	68.0%		952		0		0		Aug 6 2021
	A2813sonj3F_PREMIX_Plate_Plate 01_B04.ab1	72.6%		944		0		0		Aug 6 2021
	A2813sonj5R_PREMIX_Plate_Plate 01_A05.ab1	59.1%		908		0		0		Aug 6 2021
	A2813sonj9R_PREMIX_Plate_Plate 01_H05.ab1	61.3%		666		0		0		Aug 6 2021
	A2813sonjA2_PREMIX_Plate_Plate 01_C05.ab1	42.2%		665		0		0		Aug 6 2021
	A2813sonjBi_PREMIX_Plate_Plate 01_E03.ab1	74.8%		912		0		0		Aug 6 2021
A4846	A4846 Assembly Contig arL/BrH	98.7%	100.0%	564	530	0	0	0	0	
	A4846sonjarL_PREMIX_Plate_Plate 01_B03.ab1	93.4%		541		0		0		Aug 6 2021
	A4846sonjBrH_PREMIX_Plate_Plate 01_D03.ab1	87.0%		544		0		0		Aug 6 2021
	A4846 Assembly Contig 18S	98.4%	100.0%	1,828	1,782	0	0	1	1	
	A4846sonj1F_PREMIX_Plate_Plate 01_G07.ab1	97.2%		943		0		0		Aug 6 2021
	A4846sonj3F_PREMIX_Plate_Plate 01_C04.ab1	96.4%		945		0		0		Aug 6 2021
	A4846sonj5R_PREMIX_Plate_Plate 01_H07.ab1	96.9%		946		0		0		Aug 6 2021
	A4846sonj9R_PREMIX_Plate_Plate 01_A06.ab1	97.5%		660		0		0		Aug 6 2021
	A4846sonjA2_PREMIX_Plate_Plate 01_D05.ab1	92.4%		665		0		0		Aug 6 2021
	A4846sonjBi_PREMIX_Plate_Plate 01_F03.ab1	97.8%		942		0		0		Aug 6 2021
A5771	A5771 Assembly Contig arL/BrH	92.7%	99.6%	562		0		0		
	A5771sonjarL_PREMIX_Plate_Plate 01_A07.ab1	93.8%		536		0		0		Aug 6 2021
	A5771sonjBrH_PREMIX_Plate_Plate 01_D07.ab1	75.8%		545		0		0		Aug 6 2021
	A5771 Assembly Contig 18S	98.6%	99.9%	1,847	1,804	0	0	12	12	
	A5771sonjBi_PREMIX_Plate_Plate 01_G03.ab1	91.1%		948		0		0		Aug 27 2021

8 In Geneious, click on an assembly. In the spreadsheet, note what the sample ID is, the assembly or file name, and then scroll to the right to see what the HQ% is and write that down ("HQ% Before"), the next column over should have the sequence length and note that ("Sequence Length Before"), and 13 columns to the right should be "Ambiguities" ("Ambiguities Before"), so note that number, too.



Geneious doesn't tell you how many gaps there are, but you can find out for yourself, by clicking the two blue arrows that point towards each other, and then scrolling through the consensus sequence - it's at the top and highlighted in blue - to find gaps. They are highlighted in red and are marked with a dash. Count them, and note that number, as well.



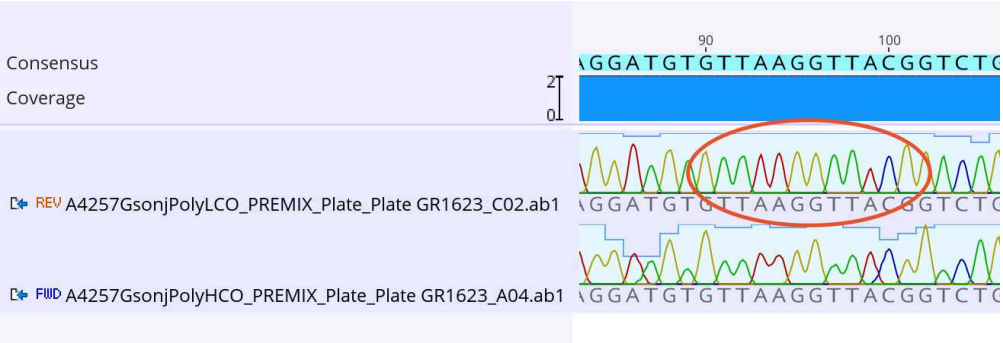


The red highlighted R is an ambiguity, it means Geneious can't quite reliably make a call on which nucleotide should be in that space.

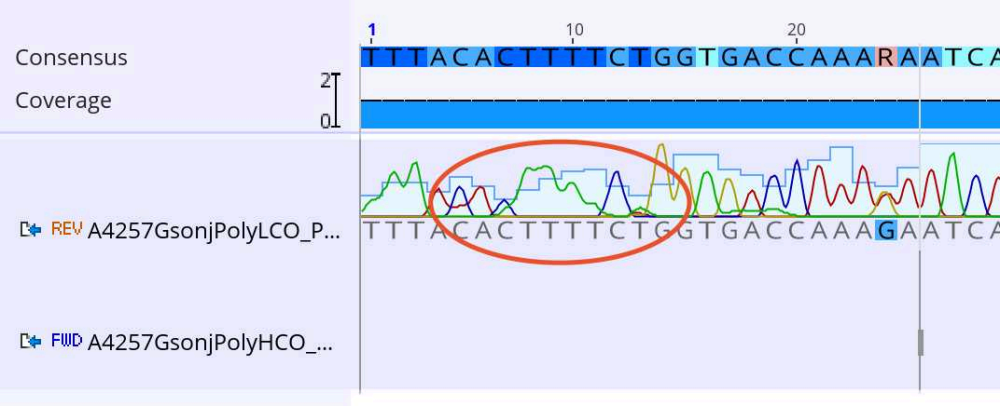
9 Once you are done noting the "Before" stats for the assemblies, it's time to edit them, so their HQ% goes up as close to 100% as you can get it. But don't worry if it's not exactly 100%! Sometimes it's better to have a longer sequence and a lower HQ%, especially if you look at the data and you trust it.

9.1 Look at the front of the assembly. Light blue is very trusted (both the forward and reverse raw read, or just one of them if it's longer than the other as long as it's clear) support the location of that nucleotide. The darker the blue, the less trusted that data is.

You can see how much confidence there is in the placement of a certain nucleotide by the colorful lines above the bottom two raw reads. In some cases, you can see clear peaks:



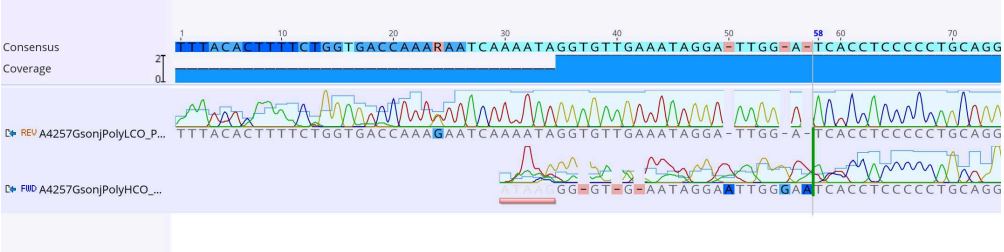
And in other cases, it's really messy:



The clearer and taller the peak, the better.

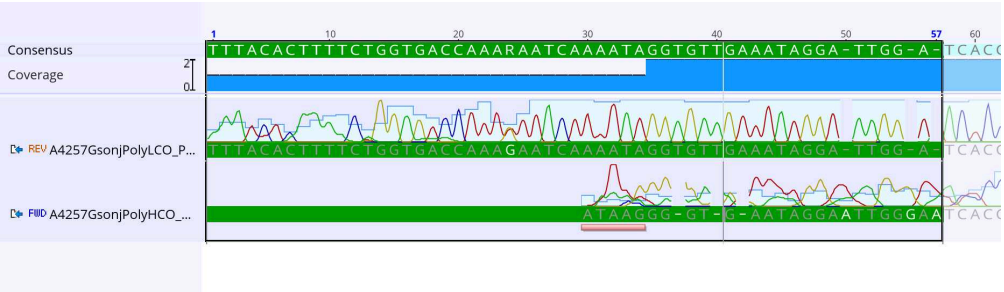
9.2 Decide where you want to cut away from the front. You CANNOT cut away parts from the middle, you have to cut everything from the beginning up until a certain point at the front. You want to get rid of as much dark and medium blue as possible, and also get rid of ambiguities if possible without sacrificing too much length (i.e. you don't want to have only 100 base pairs of a gene that you could usually get 700 base pairs from - in that case, it's better to leave an ambiguity or a medium blue somewhere). It's not as essential to also get rid of gaps, but if you can, it's usually advisable.

In my case, I'm choosing to get rid of everything from the beginning including the three gaps (to where the green vertical line is):

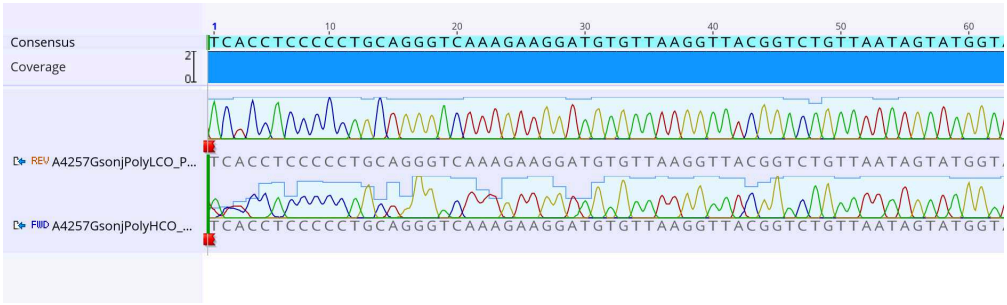


You don't have to, but for my personal records, I like to put my cursor up to where I intend to cut away the less trustworthy assembly, and take a screen shot. Geneious does keep a record of what was done, but sometimes it's just easier to go back and look at a screenshot that's kept in an organized way to know what happened, in case you ever need to check.

9.3 Click in the consensus sequence to after the nucleotide you want to cut, and then drag the mouse to the beginning of the alignment but at the bottom left corner - so you make sure two (or six if you're editing 18S) raw reads AND the consensus sequence are highlighted:

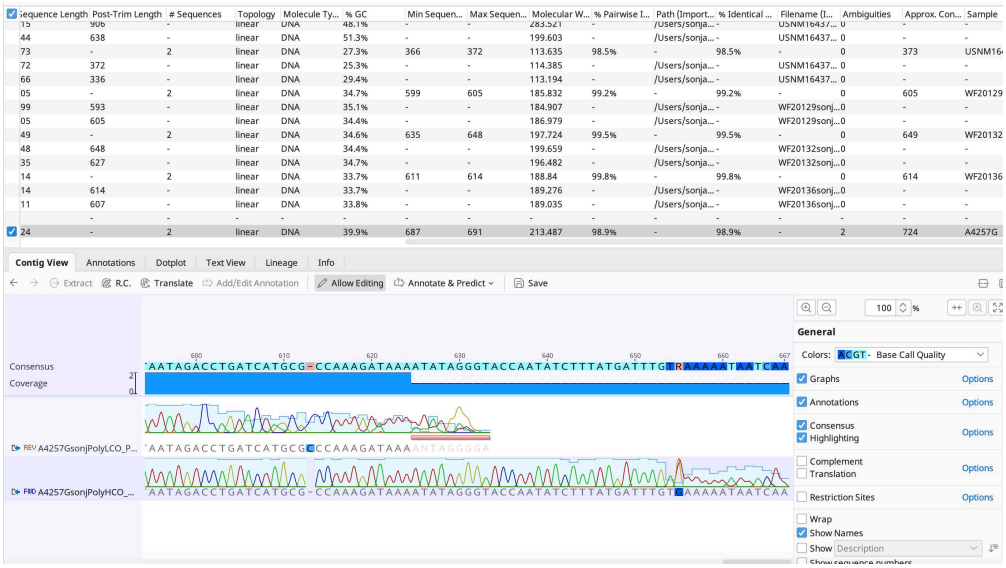


9.4 Click your "delete" button! If a pop-up window asks you to allow changes, yes, allow changes. Your sequence should look like this:

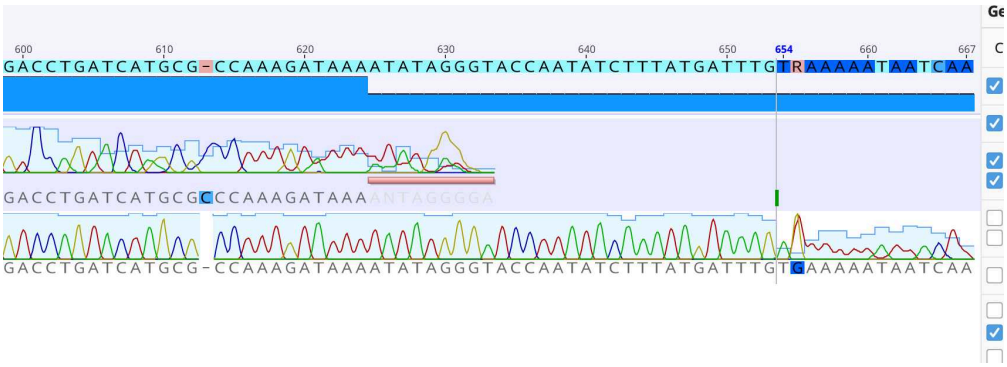


If you hover over the red tags, it will tell you what you deleted.

9.5 Then, look at the end of your assembly, and decide from where onward you want to cut the consensus sequence. Again, you CANNOT cut away parts from the middle, you have to cut everything from the nucleotide you want to cut from all the way to the end of the sequence, or it will interfere with your phylogenetic analyses.

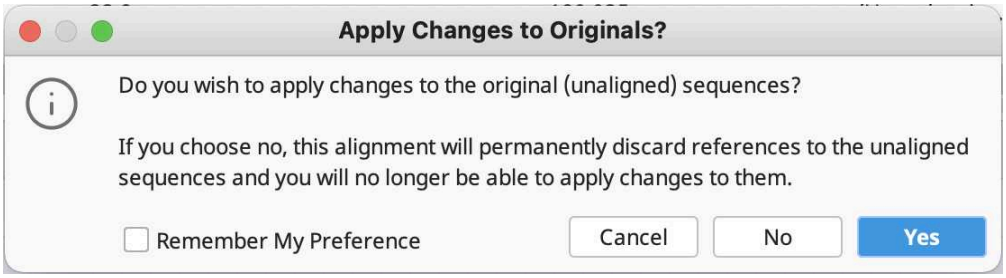


In this case, I am choosing to cut from nucleotide 654:



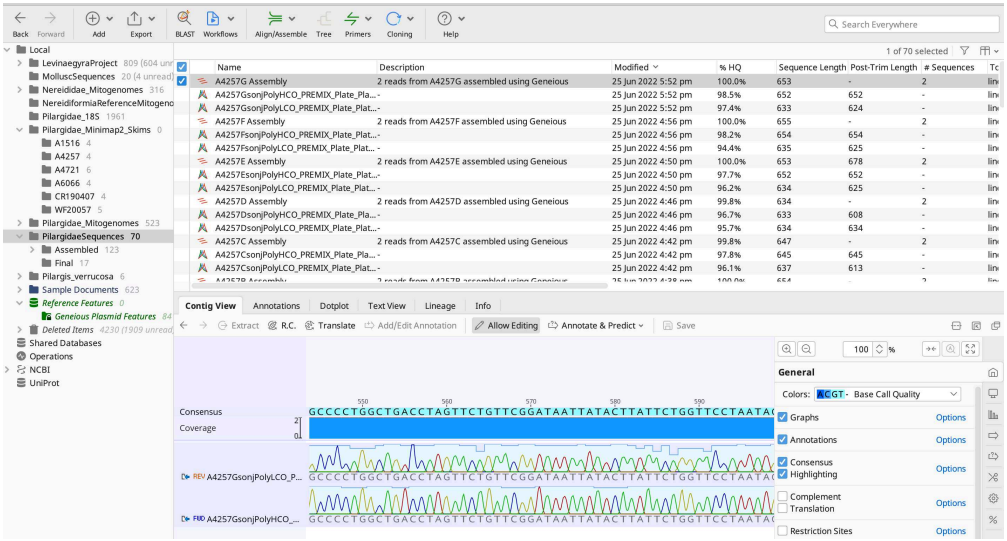
This is because the support for the rest of the spaces after the gap seems pretty good to me. I could however, just as well delete everything including the gap onward. Your call.

9.6 After deleting both ends, press "command + S" to save the changes. This popup will appear:



Select "Yes".

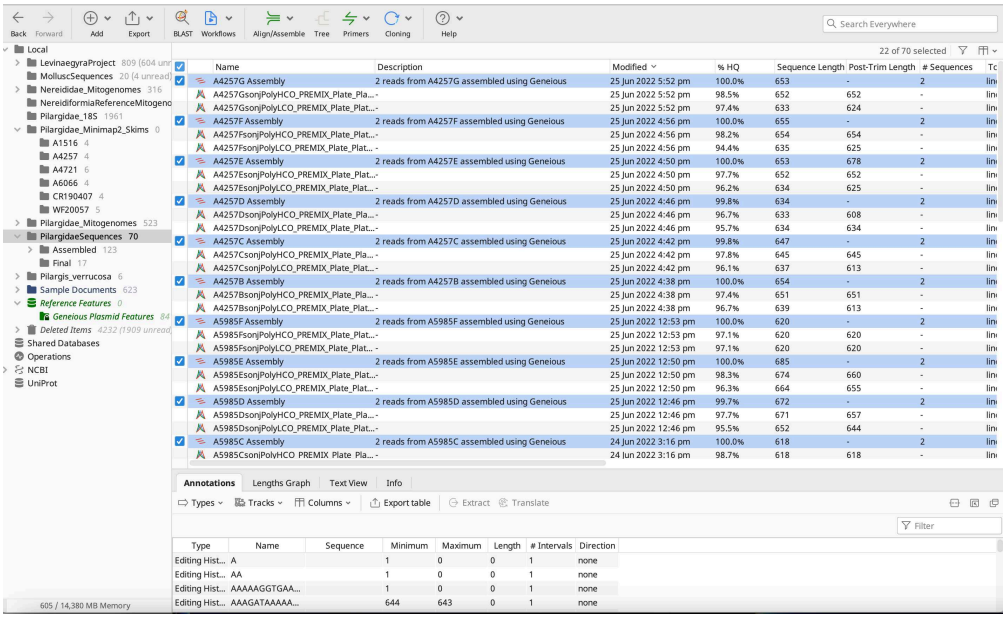
If you have Geneious set to sort files by "Modified" (which I recommend), this will put the just modified assembly and its corresponding raw reads at the top of your file list. It will look like this:



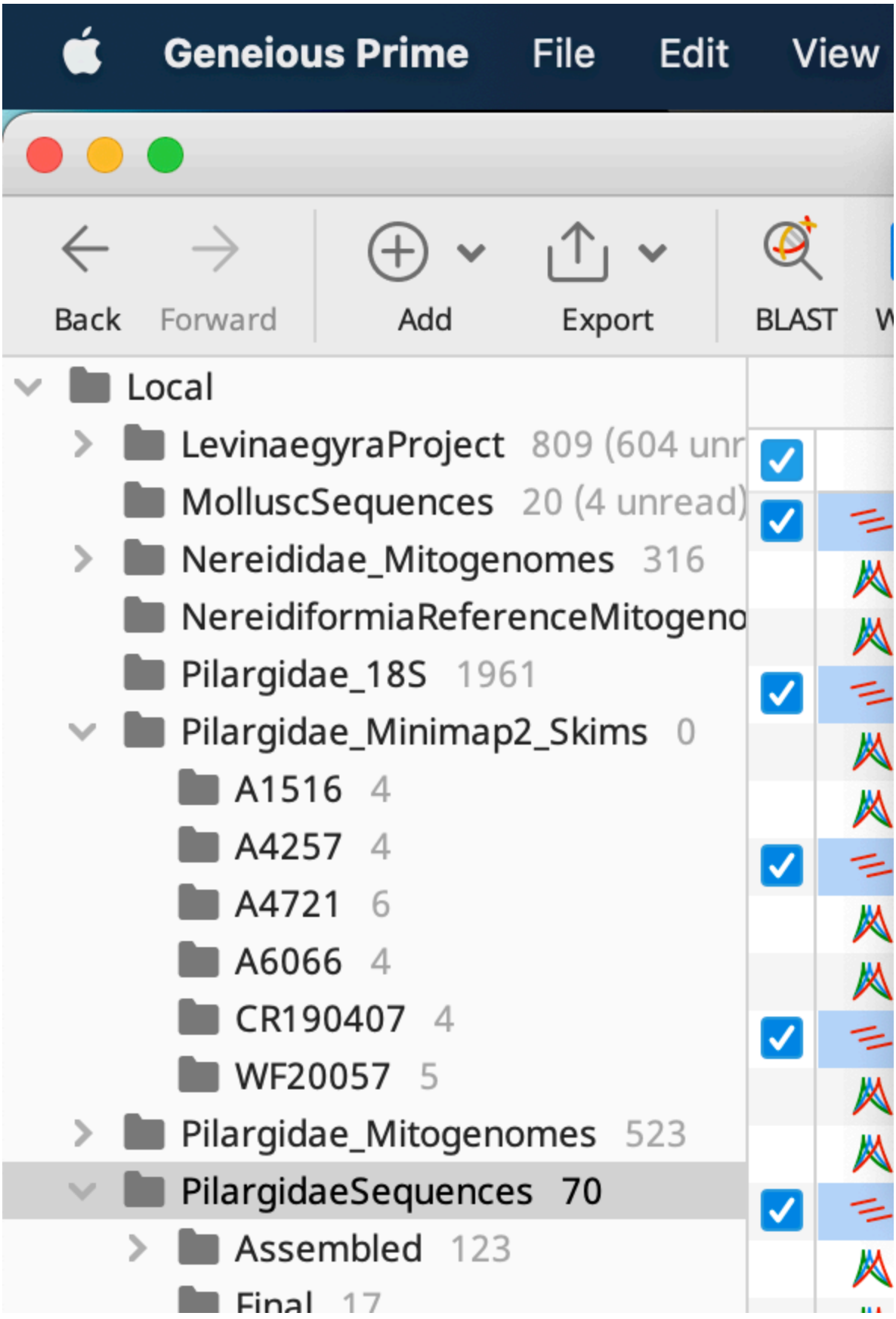
10 Look at your edited assembly, and note what the HQ% is now after you've edited, what the Sequence Length is, how many ambiguities there are, and count how many gaps there are, like before, and write this down in your excel spreadsheet file.

11 Repeat from step 8 to 10 (noting down stats and editing) for every assembly you have.

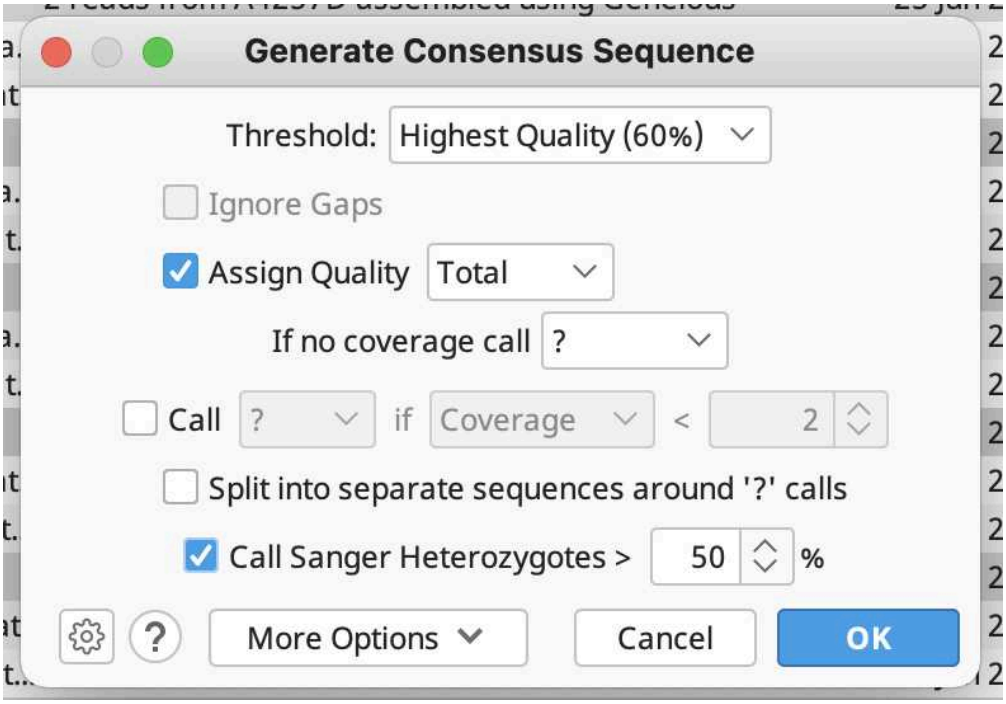
12 Once you're done, select all of the edited assemblies in Geneious.



13 Then in the tool bar, click on "Tools" and then "Generate Consensus Sequence..."

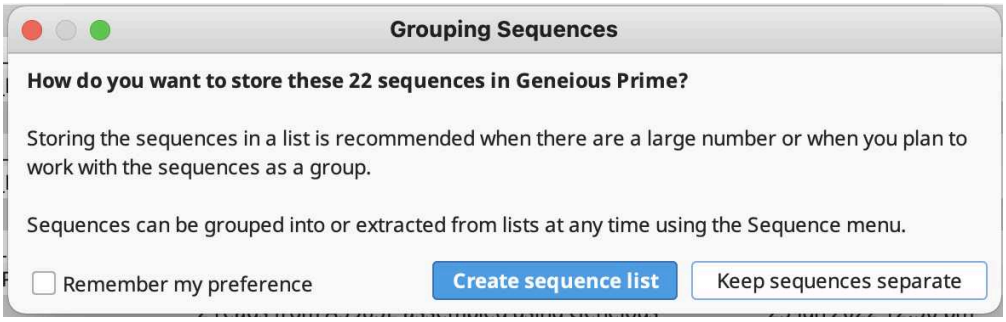

Geneious Prime
File
Edit
View
Back
Forward
Add
Export
BLAST
Local
LevinaegyraProject 809 (604 unr
MolluscSequences 20 (4 unread)
Nereididae_Mitogenomes 316
NereidiformiaReferenceMitogeno
Pilargidae_18S 1961
Pilargidae_Minimap2_Skims 0
A1516 4
A4257 4
A4721 6
A6066 4
CR190407 4
WF20057 5
Pilargidae_Mitogenomes 523
PilargidaeSequences 70
Assembled 123
Final 17

14 You will get this popup:

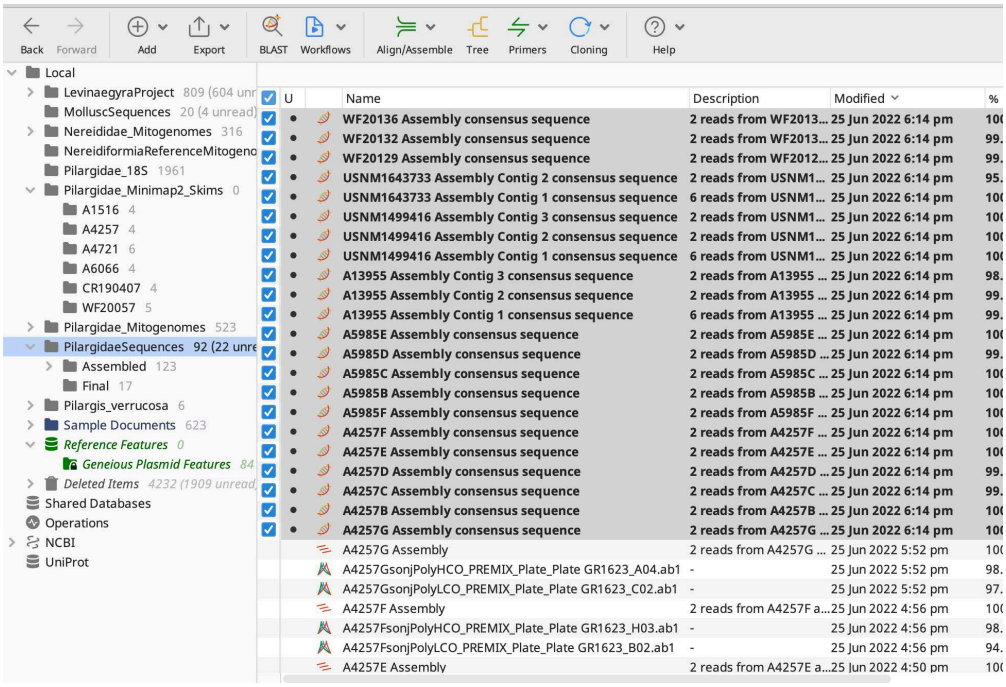


Generally, everything should be left as default, but check to make sure yours looks like this, too. Click "OK".

15 With this popup, choose "Keep sequences separate".



16 The folder in Geneious that you are working in should look something like this:

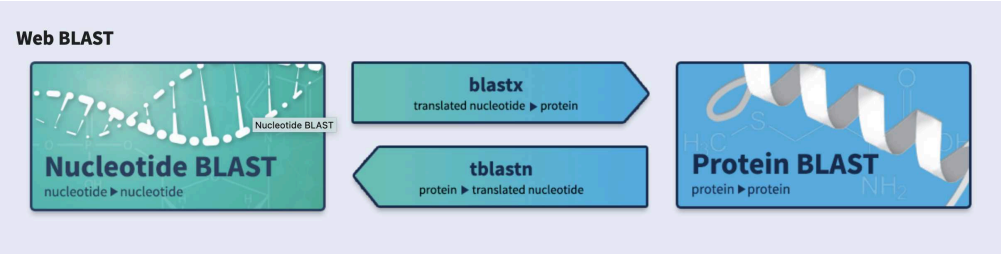


Drag the these sequences into the "Final" folder you created at the beginning, and drag the rest of the assemblies and raw reads into the "Assembled" folder you also created at the beginning. You can choose to have extra folders within the "Assembled" folder for each of the specimens you sequence, to aid in extra organization.

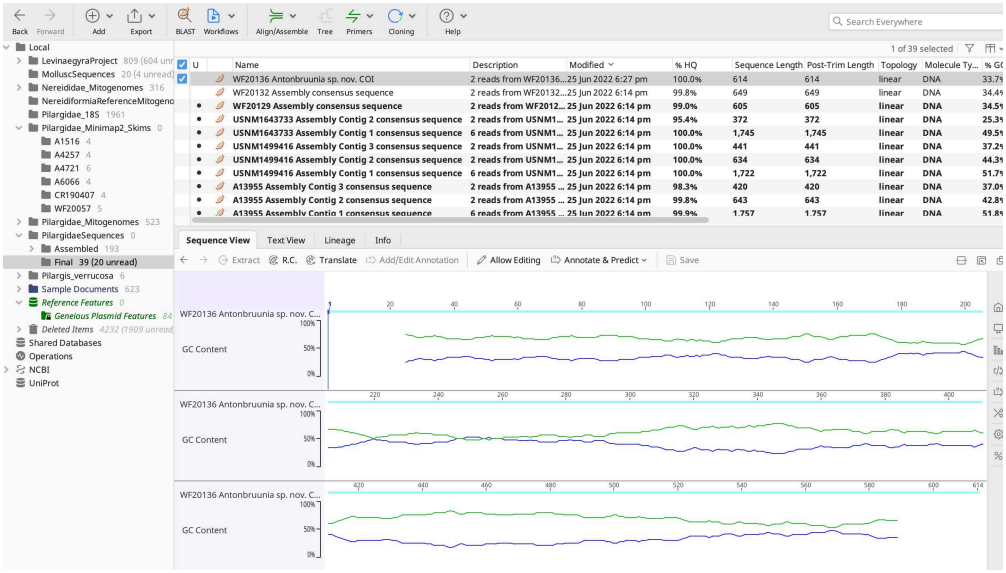
This should leave your working folder empty for the next time you'll De Novo assemble raw reads from Eurofins.

17 Blast: **This next step is VERY IMPORTANT even if it seems like you're done**. Now it's time to blast the sequences, a) to make sure they're what they're supposed to be, and not some weird contamination either on our part or Eurofins' part, and b) to make sure your assembly does not include errors (predicted stop codons or frameshifts). ***These errors are easily fixable now, so do not risk having to redo your analyses and rework your thesis/paper.***
Even if the contamination check is ok, **skipping the quality control step can come back to bite you during the GenBank submission step at the end of your project**. GenBank will automatically block your entire submission if it detects errors in any sequence. Do this now, while your .ab1 files and Geneious are easily accessible. Not years later when you have lost track of them and may not have easy access to Geneious. Reproducibility is critical in science, so the sequences you put on GenBank should reflect what you actually used in your analyses.

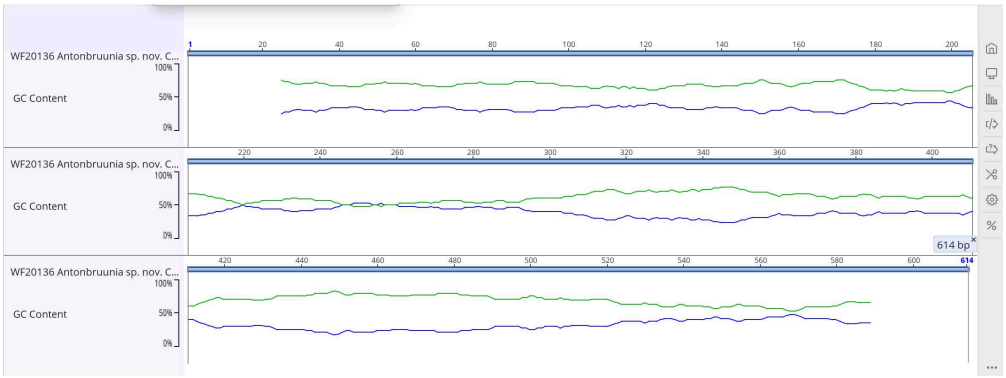
17.1 Go to https://blast.ncbi.nlm.nih.gov/Blast.cgi and click on "Nucleotide BLAST" (on the left).



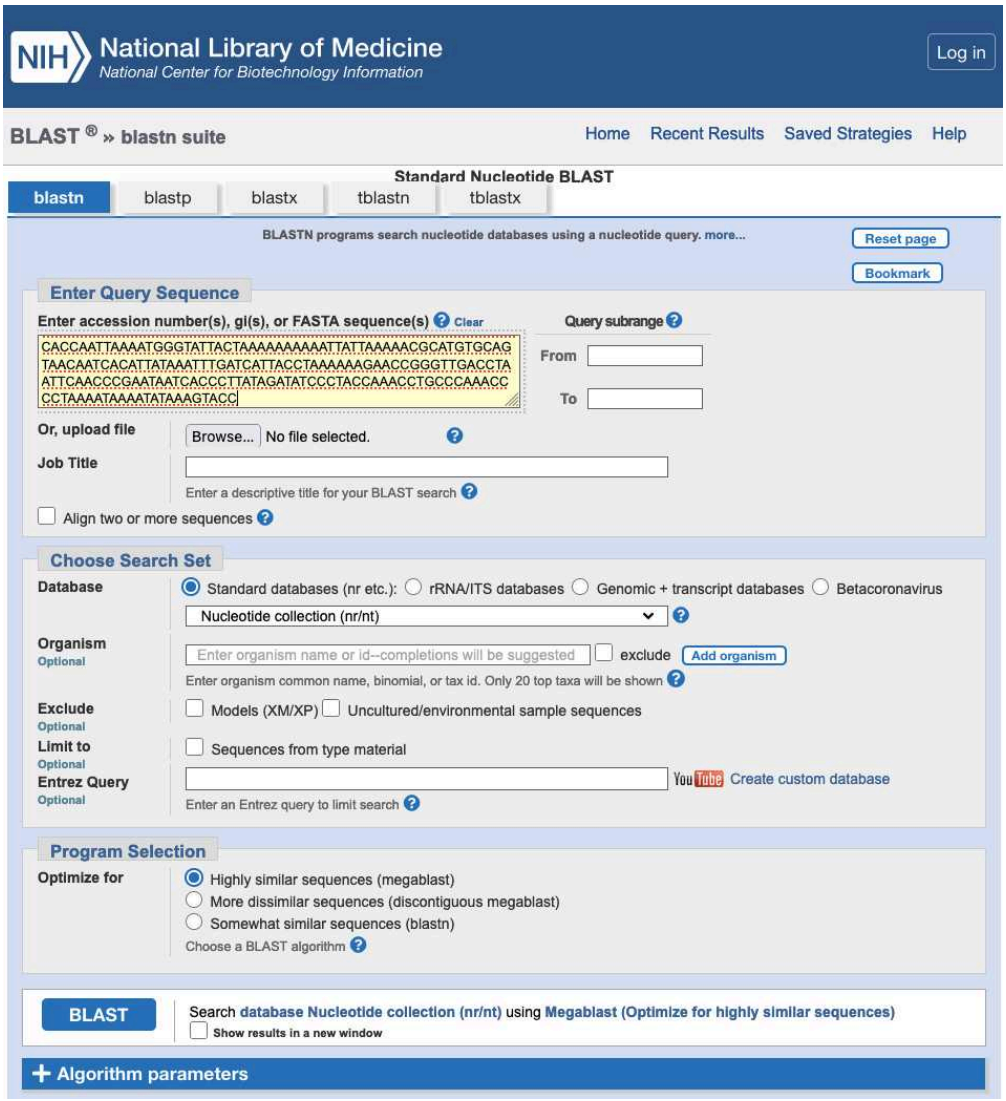
17.2 Go to the "Final" folder in Geneious where you just dragged the generated sequences. You can rename them to something that tells you more about what they actually are (i.e. add a "COI" or "18S" at the end of whatever is there already in the name, or just keep the sample ID and add what gene it is, whatever helps you know what it is). Select the first one, and it should look something like this:



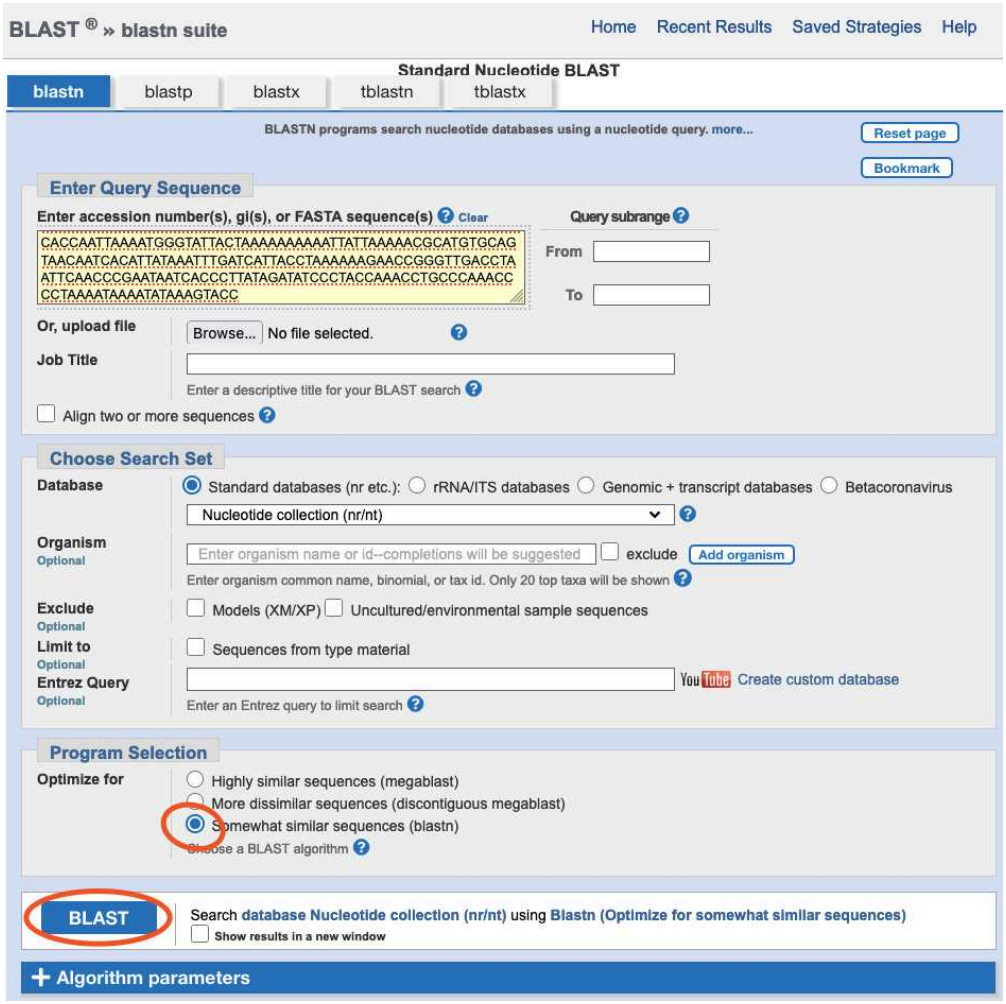
17.3 Click at the beginning of the light blue sequence, and drag to the end of it, so that it turns a darker blue (it's highlighted, selected). Then click "command + C" to copy the sequence.



17.4 Go back to the NCBI blastn web page, and paste the sequence into the empty box in the "Enter Query Sequence" section.



18 Select "Somewhat similar sequences (blastn)" under the "Program Selection", and then press "BLAST".



19 Check that the overall results make sense given the known identification of your sequence (e.g., not bacterial contamination).

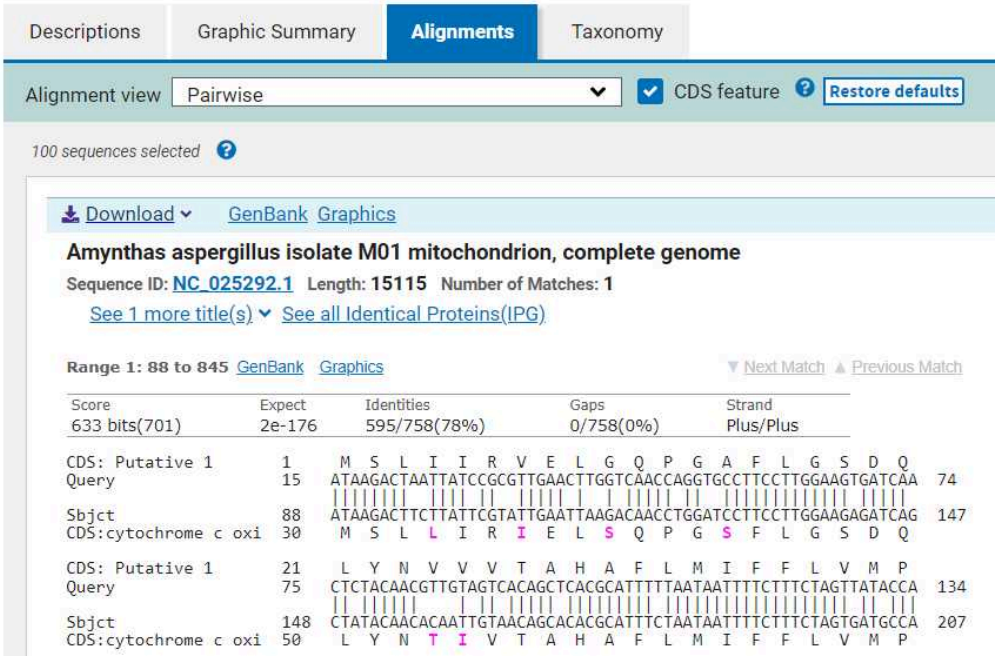
20 Click on the 'Description' link for the the top GenBank record that your sequence aligns with.

Under the 'Strand' heading, it should say ‘Plus/Plus’. This indicates that your sequence is oriented in the correct forward 5’ direction. If it says ‘Plus/Minus’, then reverse complement the alignment generated in step 1.

21 **Additional quality control steps for protein-coding genes such as COI**

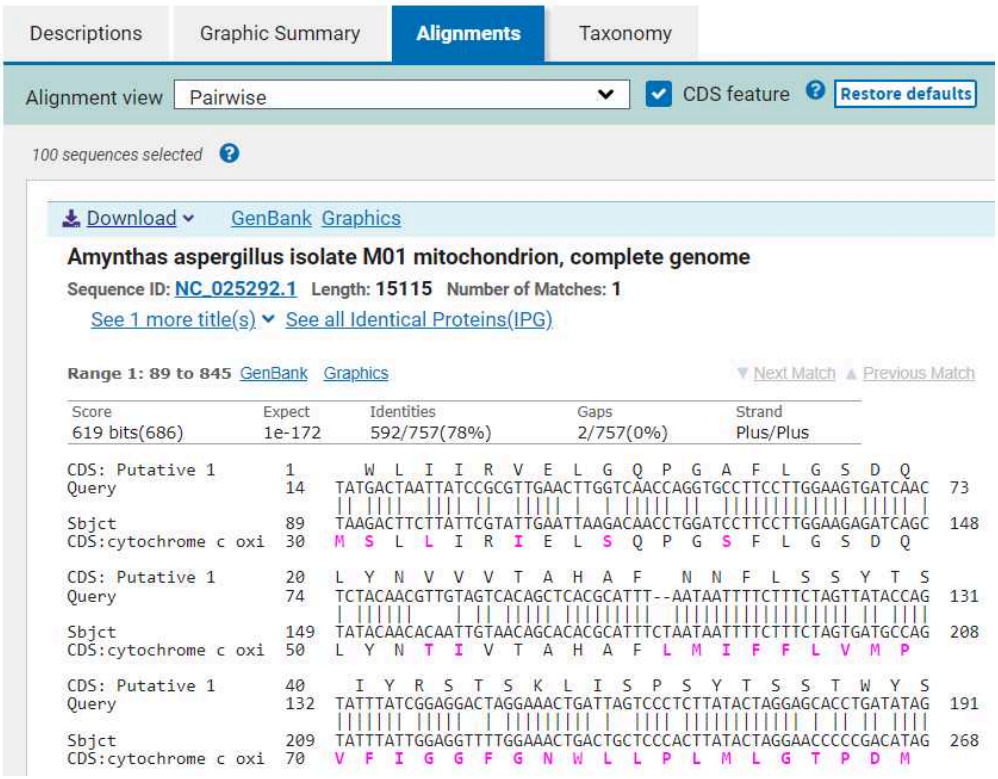
21.1 Check the box for 'CDS feature' to display the expected amino acids for your sequence compared to those of the BLAST hit. Mismatches will be shown in pink. Overall the

majority of amino acids should match. Some variation is expected, of course.



Example BLAST alignment showing an acceptable CDS

If you see a long string of mismatches, there is likely a frameshift in your sequence. A nonzero number of gaps would also suggest a frameshift.



Example BLAST alignment showing a frameshift

If you see a frameshift, revisit that region of the sequence in your .ab1 files in Geneious and check if any base calls need to be edited.

21.2 Also check that there are no stop codons (marked by *) in your coding sequence.

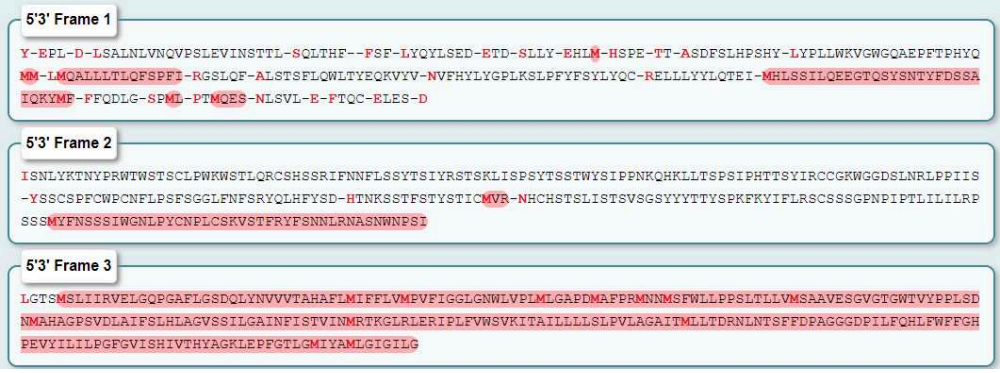
Example BLAST alignment showing a frameshift (pink text) and stop codon (*)

Revisit your .ab1 files if needed.

21.3 Now check the rest of your sequence for any remaining frameshifts and stop codons. (The BLAST step above was important to check for contamination and tell you the correct reading frame, but it only checked the parts of your sequence that were aligned with BLAST hits.)

For many sequences you may wish to use Mesquite, but if you are only working with a few sequences, you might find it easiest to use a reputable online translation tool such as Expasy (**https://web.expasy.org/translate/**). Choose the appropriate genetic code for your organism, such as "invertebrate mitochondrial". Make sure that one of the 5'3' (forward) reading frames produces an uninterrupted amino acid sequence with no stop

codons (-). This amino acid sequence should overlap with the CDS result from your sequence during the BLAST check.



Example translation in Expasy

Revisit your .ab1 files as needed to resolve stop codons or frameshifts.