Jun 12, 2025 Version 1

NCBI COI Submission V.4 V.1

Forked from a private protocol

DOI
https://dx.doi.org/10.17504/protocols.io.e6nvw1xzdlmk/v1

Avery S Hiley[1], Dakota Betz[2]

[1]UCSD- Scripps Institution of Oceanography; [2]ucsd

Avery S Hiley: seahatch3@gmail.com | ahiley@ucsd.edu | http://www.spineless.info/avery-hiley.html;

Rouse Lab

Charlotte A Seid
University of California, San Diego

DOI: https://dx.doi.org/10.17504/protocols.io.e6nvw1xzdlmk/v1

Protocol Citation: Avery S Hiley, Dakota Betz 2025. NCBI COI Submission V.4. **protocols.io** https://dx.doi.org/10.17504/protocols.io.e6nvw1xzdlmk/v1Version created by **Dakota Betz**

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

Created: October 22, 2024

Last Modified: June 12, 2025

Protocol Integer ID: 110461

Keywords: uploading mitochondrial coi sequence, mitochondrial coi sequences to genbank, ncbi cois, sequences to genbank, genbank, ncbi

Disclaimer

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

Abstract

Protocol for uploading mitochondrial COI sequences to GenBank

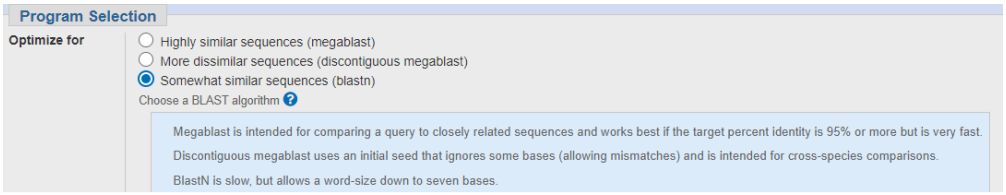
Before start

It is important to check the quality of your sequences (using the instructions below to recognize frameshifts, stop codons, and contamination/mislabeling) before submitting them to GenBank. GenBank's automated error detection will flag suspected frameshifts and stop codons within minutes of your submission. If justified, you can email GenBank to override these error reports, but note that a single error will stop your entire submission from progressing.

Quality Checking and Submission of COI Sequences to GenBank

1 First and foremost, gather all your COI sequences that need to be uploaded and align them in **Geneious**, **Mesquite**, or using the online **MAFFT version 7** server: https://mafft.cbrc.jp/alignment/server/.

2 Copy and paste one of your COI sequences into the 'Enter Query Sequence' field on NCBI's **Standard Nucleotide BLAST** platform: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome. You may need to adjust your choice of BLAST algorithm (see three options below) if your sequence is very divergent from what is on GenBank.

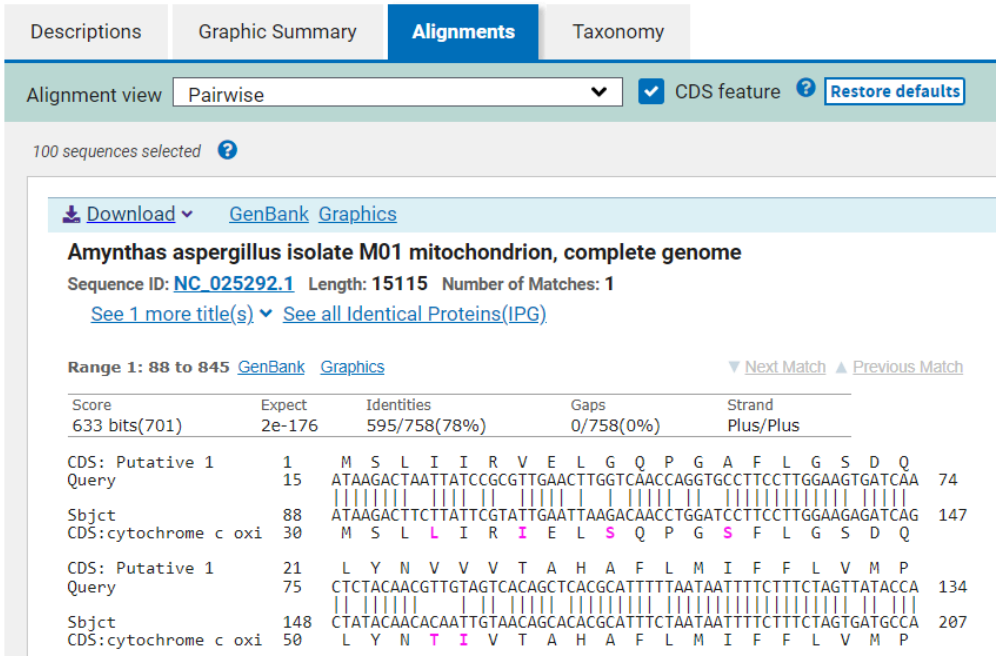


Three BLAST algorithm options (as of 2024-10-17)

2.1 Check that the overall results make sense given the known identification of your sequence (e.g., not bacterial contamination). Click on the 'Description' link for the the top GenBank record that your sequence aligns with.

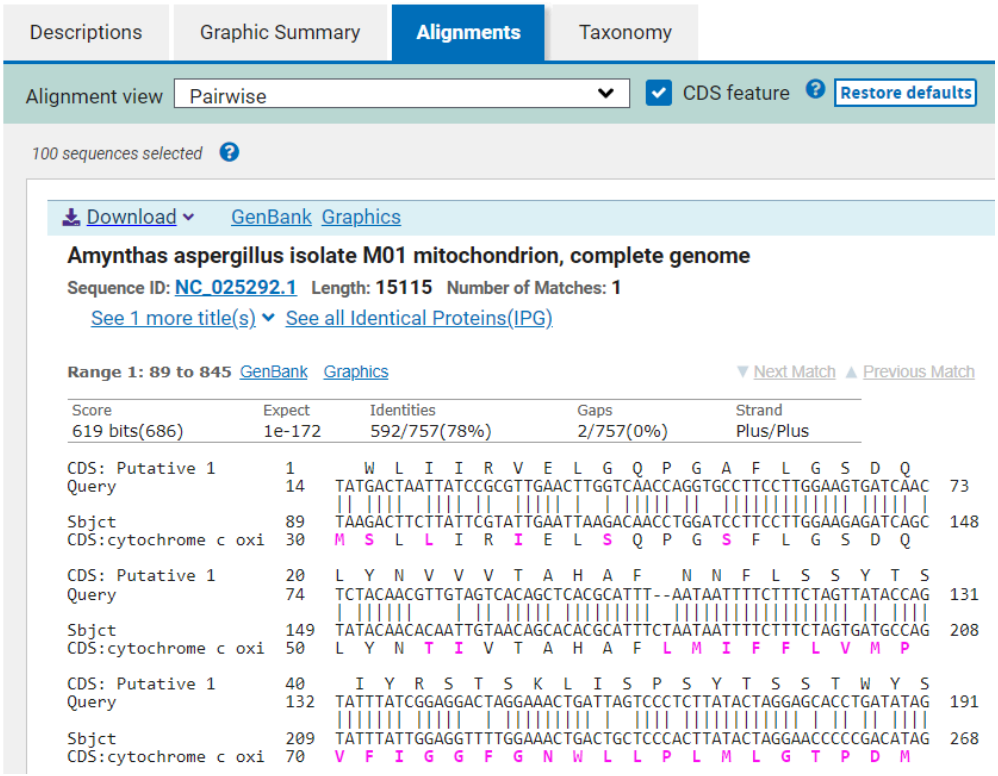
2.2 Under the 'Strand' heading, it should say ‘Plus/Plus’. This indicates that your COI sequence is oriented in the correct forward 5’ direction. If it says ‘Plus/Minus’, then reverse complement the COI alignment generated in step 1.

2.3 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.



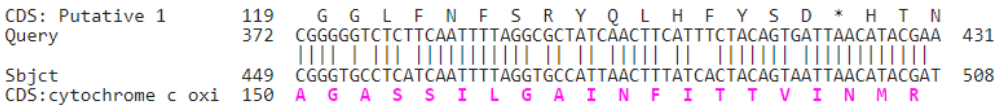
Example BLAST alignment showing an acceptable CDS

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

2.4 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 (*)

3 Please make sure to avoid ambiguities in your sequences. GenBank does accept them, though. If a COI sequence has lots of bases that aren't ACGT (e.g. N, ?, K, M, other IUPAC codes, etc.):

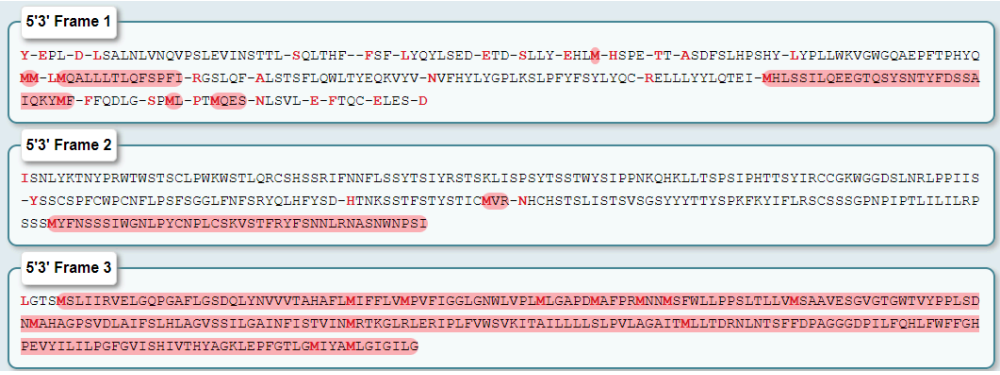
3.1 Please look at your original De Novo assembly in **Geneious** and check if you can make a nucleotide base call. Compare the amplitudes of the Forward/Reverse peaks at the corresponding nucleotide positions in the chromatogram (.ab1 files), and choose the nucleotide with the greater amplitude.

3.2 Check if the ends of the sequence assembly need to be trimmed due to low/poor quality.

4 Make any nucleotide edits and/or deletions as necessary before proceeding.

5 Check your entire sequence for frameshifts and stop codons. (Note that the BLAST step above only checked the parts of your sequence that aligned with BLAST hits.)

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 (-).



Example translation in Expasy

If you are working with an alignment of many sequences, see the steps below.

5.1 Open your edited COI alignment in **Mesquite**, and click on the colorful rubix cube icon on the left that says 'Character Matrix'. Then click 'List & Manage Characters'.

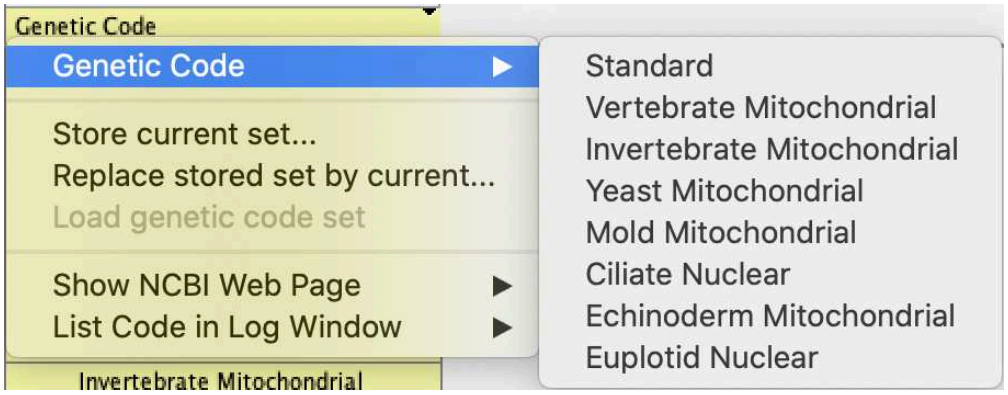
5.2 Highlight/select all the columns and rows. Click the 'Codon Position' column heading > 'Set Codon Position' > 'Minimize Stop Codons'.

5.3 In the top task bar, click 'Columns' > 'Current Genetic Codes'.

5.4 Click the 'Genetic Code' column heading > 'Invertebrate Mitochondrial'.

Note

If your COI sequences pertain to a different genetic code, you may choose one of the following alternative options in **Mesquite**:



5.5 In the top task bar, click ‘Characters’ > ‘Make New Matrix from’ > ‘Translate DNA to Protein’.

5.6 View the protein translation of your character matrix, and make sure there are no black stop codons with an asterisk.

5.7 If the translation looks good (no stop codons), proceed to view the original nucleotide character matrix again. Export the final COI alignment as a fasta (DNA/RNA) file.

Note

Do not check 'include gaps' when you export the COI alignment as a fasta file in **Mesquite**.

6 Open this fasta file in **TextWrangler** or **BBEdit**. Edit all of your sequence headers to match this format:

>SIO:BIC:A9919 organism=Peinaleopolynoe mineoi Peinaleopolynoe mineoi voucher SIO:BIC:A9919 cytochrome c oxidase subunit I (COI) gene, partial cds; mitochondrial

If you have a lot of sequences, you might find it efficient to work in **Excel** and use a concatenation formula to generate this header for you.

6.1 In this example, “SIO:BIC:A9919” is the sequence ID for the specific COI sequence. Please use either the SIO:BIC (SIO:BIC for Scripps Institution of Oceanography, Benthic Invertebrate Collection) catalog number or a different institution’s catalog number here to represent the sequence ID. If the specimen is deposited at a different institution, the catalog number will instead be preceded by that specific institution's abbreviations. A couple examples of alternative catalog numbers are listed below:

- Muséum national d'Histoire naturelle (MNHN): MNHN:IA:2010-399
- Museum of Comparative Zoology (MCZ): MCZ:70173

Note

Do not include any spaces in the sequence ID. If necessary, use a colon as a replacement for a space.

6.2 “[organism=Peinaleopolynoe mineoi]” is obviously where you insert the species name for the corresponding sequence ID.

Note

You **do not** need to create placeholder names for new species that are not yet published. In this example, *Peinaleopolynoe mineoi* was a new species that was not yet published during the time of its GenBank COI submission.

The GenBank staff will automatically create placeholder names for new species included in your submission. Once your manuscript is officially published, email **gb-admin@ncbi.nlm.nih.gov** (with the corresponding submission ID as the subject line) and request any changes that need to be made, including but not limited to reverting the placeholder names back to the full new species names and updating the publication details (e.g. journal, article title, authors, publication date, doi link, etc.).

6.3 "Peinaleopolynoe mineoi": Repeat the species name here.

6.4 "voucher SIO:BIC:A9919": Follow this proper format to identify the corresponding SIO:BIC voucher with our institution's abbreviations. As mentioned in step 12.1, the voucher should be identified with the abbreviations of whichever institution it is deposited at, followed by the catalog number for that specimen.

Note

The voucher description should be the exact same as your sequence ID from step 12.1.

6.5 "cytochrome c oxidase subunit I (COI) gene, partial cds; mitochondrial" should be listed at the end of the aforementioned information.

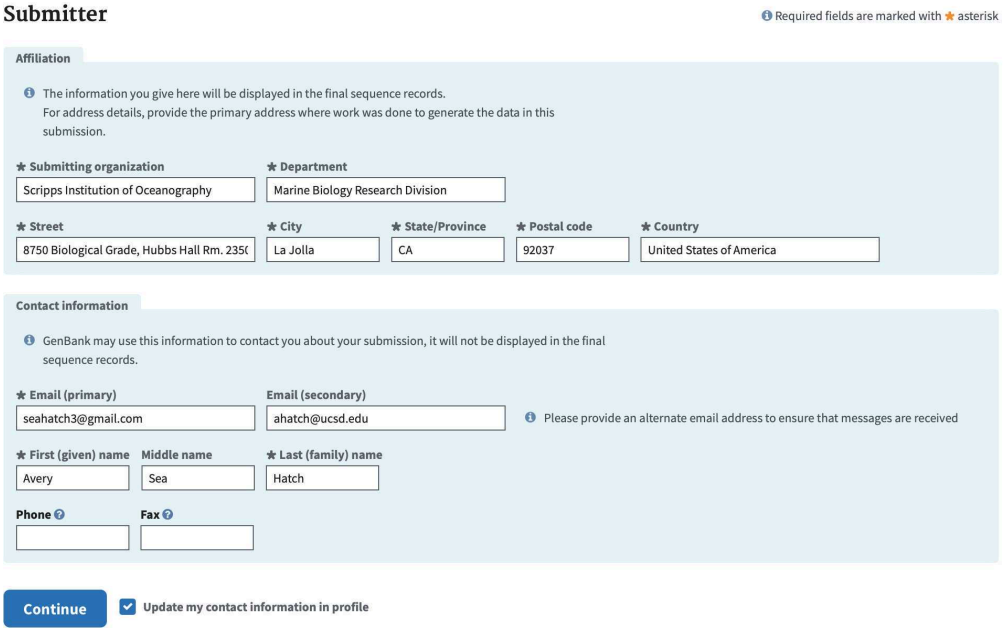
7 After all of your sequence headers follow this format, save the edited fasta file.

8 Create an online account for NCBI's **Submission Portal** platform: **https://submit.ncbi.nlm.nih.gov/**.

9 Start a new submission.

10 *Submission Type*: 'Metazoan (multicellular animal) Mitochondrial COX1'.

11 *Submitter*: Fill out the corresponding information (see screenshot below for our lab details). Make sure to check 'Update my contact information in profile' in order for future submissions to use this specific info by default.



12 *Sequencing Technology*: Select 'Sanger dideoxy sequencing' and 'Assembled sequences (each sequence was assembled from two or more overlapping sequence reads)'.

In rare cases, you might submit a COI sequence that was assembled from genome or transcriptome data, rather than Sanger sequencing. Choose the appropriate method (e.g., Illumina) and assembly program with mandatory version number or date (e.g., Agalma #.#.#).

13 *Sequences*: Select 'Release on specified date or upon publication, whichever is first' or release immediately if you are late in uploading your sequences (this should almost never be the case). Typically you will want to choose a year in advance to be safe. Next, upload your COI fasta file that was completed in step 7.

14 *Source Info*: Under 'Do your sequence IDs represent one of these?', select 'Specimen-Voucher' if you correctly followed the headers format in step 6.

15 *Source Modifiers*: This is the section where you add key details that you would like to be attached to the sequences (e.g. locality and depth). However, only the organism name and specimen-voucher are required for submission. You should have already included these in the previous steps, so you may continue if you do not wish to apply additional source modifiers.

Note

Since all the specifics should be in your publication, you may typically stick with locality (column name = Geo_Loc_Name) and depth (column name = Altitude). The locality information in the 'Geo_Loc_Name' column [formerly 'Country'] must go from broad to specific. For example, "Mexico: Gulf of California, Pescadero Basin" follows this format. The depths entered in the 'Altitude' column must be negative (e.g. "-3676 m."). A list of all the other available source modifiers and their descriptions, including the correct formats to use for their responses, may be found at the following link: https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers.

Required columns:
- Sequence_ID [the name of the sequence in your fasta file, so 'SIO:BIC:A9919' in the previous example; for our purposes, this is typically the same as Specimen_voucher]
- Specimen_voucher
- Organism

Recommended by our lab:
- Bioproject [use PRJNA1136884 for SIO-BIC sequences unless you already have another Bioproject set up; do not use this for non-BIC sequences]

Important if your sequences did not come from SIO-BIC specimens:
(Not strictly necessary if your sequences are from SIO-BIC specimens, because these details are available through SIO-BIC.)
- Geo_Loc_Name [formatted from broad to specific, for example: Pacific Ocean:East Pacific Rise]
- Lat_Lon [format as decimal degrees with N/S, E/W indicated, for example: 4.9871 N 87.4433 W]
- Altitude [use a minus sign for ocean depths, for example: -243 m]
- Collection_date [format like so: 20-Jan-2019]
- Collected_by

If you would like to add a couple of key source modifiers, then choose one of the following options under 'How do you want to apply source modifiers?':

15.1 *Option 1*. 'Use a form to apply the same value for all sequences': This option is very straightforward. Choose a source modifier category on the **Submission Portal** interface and then type the response next to it. Add as many source modifiers as necessary.

Note

This option will rarely be used, unless all of your COI sequences have the exact same details for the source modifiers that you wish to add.

15.2 *Option 2*. 'Use an editable table': This is self explanatory. You can manually add columns (source modifier headers) and input the corresponding information for each sequence ID on the **Submission Portal** interface.

Note

However, please copy and paste, and save your work somewhere else for your records!

15.3 *Option 3*. 'Upload a tab-delimited table': Proceed to download the source modifier template table, or create your own. You can edit this file in several programs (**TextEdit**, **TextWrangler**, **BBEdit**, **Microsoft Excel**, or **Numbers**). Follow the specific format for each source modifier (described at the following link) applicable to your submission: https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers.

- In a text edit program, separate the information in each column by inserting 1 tab. An example of a tab-delimited table is attached herein:

  Peinaleopolynoe_COI_Source_Modi...

- If you edit the source modifier template table in **Microsoft Excel** or **Numbers**, please make sure to export the final table in the TSV (tab-separated values) or tab-separated text (.txt) file format.
- Upload your final source modifiers file to the **Submission Portal** interface.

16 *References*: Under 'Sequence authors', add yourself (unless someone else did the lab work). Select the corresponding 'Publication status' that applies to you. Add the title of your paper and 'Specify new authors' to add the names of all authors on your paper.

17 *Review & Submit*: Check the details of your submission here. If everything looks good, submit your sequences.

18 *Responding to Errors*: GenBank's automated system will email you within a few minutes if errors are detected, such as predicted frameshifts or stop codons. You must edit your submission (by uploading corrected files) or email GenBank to request an exception, or else your submission will be automatically deleted. Review the error report and troubleshoot your sequence(s).

18.1 If your sequence legitimately needs to be corrected, make the required edits and submit the new files using the "Fix" feature of the GenBank submission website. You will have to go through the entire submission form again, although many of your previous responses will have been saved.

18.2 If you extensively review your sequence and believe that GenBank's error report is incorrect, you can email them with an explanation.

For example, GenBank sometimes sends a "low confidence possible frameshift" error if your group of animals does not have many examples on GenBank and is highly divergent to other groups on GenBank. Please do check and double-check that your sequence:

- has no protein translation issues (uninterrupted 5'3' sequence of amino acids, no stop codons)
- has an acceptable alignment with a reasonably related organism on GenBank (CDS feature confirms no frameshift, no legitimate gaps)

If you are confident that you are right and GenBank is wrong, email them politely, referencing your submission number in the subject line. Do not edit your original submission. Just wait a few days. If they accept your explanation, they may not reply to your email but the submission will change status and you will receive normal GenBank accession numbers.

Note

to: gb-admin@ncbi.nlm.nih.gov
subject: GenBank Submission SUB14778704

Hello,

Thank you for the notification. I have reviewed the "low confidence possible frameshift in CDS (frame restored before end)" error in the attached COX1 sequence (SUB14778704).

From checking the translation and BLAST results of this sequence (using the CDS feature to check for frameshifts), as far as I can tell, the reading frame seems to be consistent with the best available closely related sequences on GenBank, i.e., other sipunculans such as
JN865109.1 and JN865110.1 with no gaps and 99% query cover.

Would you be willing to reconsider the "low confidence" of this reported error?

I hope it would be possible to proceed with the upload of this sequence as is, and I would appreciate any further insights.

Thank you for your input,
Charlotte