NCBI COI Submission V.2

Avery S Hiley[1]

[1]UCSD- Scripps Institution of Oceanography

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

Version 2 ▼

Aug 25, 2021

1   *Works for me*   ⤴ Share

dx.doi.org/10.17504/protocols.io.bma7k2hn

Avery Hiley
UCSD - Scripps Institution of Oceanography

ABSTRACT

Protocol for uploading mitochondrial COI sequences to GenBank

1   First and foremost, gather all your COI sequences that need to be uploaded and align them in either **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. Click on the 'Description' link for the the first GenBank record that your sequence aligns with. 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.

3   Please make sure there are no ambiguities and/or gaps in your sequences. 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   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'.

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

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

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

If your COI sequences pertain to a different genetic code, you may choose one of the following alternative options in **Mesquite**:
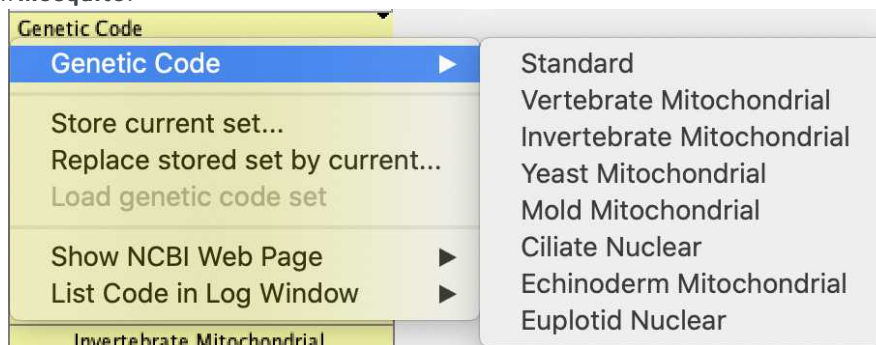
9   In the top task bar, click 'Characters' > 'Make New Matrix from' > 'Translate DNA to Protein'.

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

11   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.

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

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

12   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

12.1   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): MCZ70173

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

12.2   "[organism=Peinaleopolynoe mineoi]" is obviously where you insert the species name for the corresponding sequence ID.

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.).

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

12.4   "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.

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

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

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

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

15 Start a new submission.

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

17 *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.

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

19 *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 13.

20 *Source Info*: Under 'Do your sequence IDs represent one of these?', select 'Specimen-Voucher' if you correctly followed the headers format in step 12.

21 *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.

> Since all the specifics will be in your publication, you may typically stick with locality (column name = Country) and depth (column name = Altitude). The locality information in the 'Country' column 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.

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?':

21.1   *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.

> 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.

21.2   *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.

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

21.3   *Option 3*. 'Upload a tab-delimited table': Proceed to download the source modifier template table. 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_Modifiers.tsv**
   - 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) file format.
   - Upload your final source modifiers file to the **Submission Portal** interface.

22   *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.

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