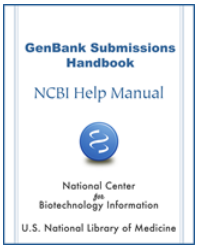
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Fetchko M, Kitts A. The "Nucleotide" Page. In: The GenBank Submissions Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



The "Nucleotide" Page

Michael Fetchko and Adrienne Kitts

Purpose

The BankIt "Nucleotide" page is where you will provide your nucleotide sequence(s), information about the molecule type, and the physical form of the sequence. The information we ask for in this section includes:

- The number of sequences you are submitting
- The actual sequence(s) you are submitting
- Whether you want the sequence available to the public on GenBank as soon as the data is processed, or if you want it released on a particular date that you provide
- If your sequence is 16S rRNA, and if it is, whether you used a chimera checking tool to test the sequence for the presence of chimeras
- The type of molecule you are submitting (i.e. is it genomic DNA, mRNA, etc.)
- Whether the sequence is linear or circular (e.g. plasmid, some viruses, cloning vectors)

The "Submission Release Date" Section

In this section tell us if you prefer your sequence to be available to the public in GenBank as soon as we process it, or if there is a particular date when you want the sequence(s) released to the public on GenBank.

This section has "Immediately After Processing" as the preset selection. If you want to set a particular date to have your sequence(s) released, click on "Release Date", and then:

- Enter the date you want into the text box using the format example you will find to the right of the release date text box.

  OR
- Place your cursor in the text box and click once. A calendar will appear (Figure 9) that you can use to select the date you want.

If you use the calendar:

Go to the month and year you want using the arrows to the left and right of the month at the top of the calendar. When you click on the day you want the sequence released, that date will appear in the release date text box in the correct format

Once the date appears in the release date text box, check to be sure you selected the correct date.

Note: the maximum time we allow for a later release of your sequence is 4 years.

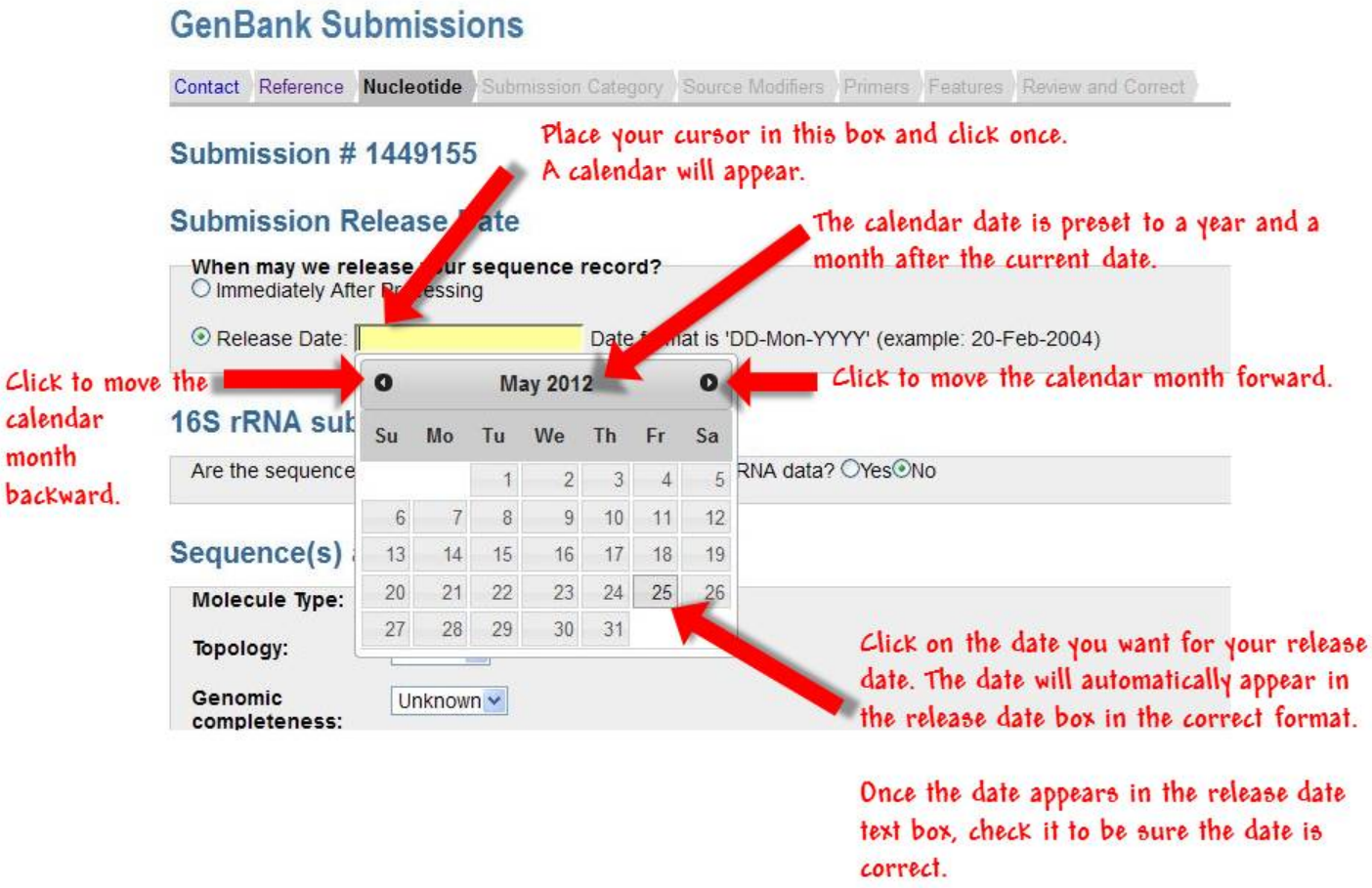


Figure 9: The Submission Release Date section of the Nucleotide page showing the calendar that appears once you place your cursor in the text box and click. Figure text provides hints for using the calendar.

The "16S rRNA Submissions" Section

In this section, tell us if you are submitting ONLY 16S rRNA sequences. If you are, tell us if you checked your sequence(s) using a chimera checking tool.

If you are submitting ONLY 16S rRNA sequence(s), click the "Yes" button " in response to the question "Are the sequences in this submission ONLY 16S ribosomal RNA data?" since the "No" response button is the preset answer to the question.

Once you tell us that you will be submitting 16S rRNA sequences, you will then be asked if you checked the sequence(s) for chimeras using a chimera checking tool. If you answered "Yes" that you did check for chimeras, you will be asked for the name of the tool you used and the version number of the tool used.

The Importance of using a Chimera Checking Tool

Removing chimeras that contaminate sequence is very important since chimeric sequence that is not removed can lead researchers using your sequence to make incorrect conclusions.

If you do not give us the name and version number of the tool you used to check your 16S rRNA sequence for chimeras, you will delay your submission, since we will contact you to get this information before any accession numbers are assigned.

Features Page Automatically Completed by BankIt when you Submit a Group of 16S rRNA Sequences

If you have a group of ONLY 16S rRNA sequences, submitting them together can save time. When you click "yes" that you are submitting ONLY 16S rRNA sequences, BankIt will automatically add 16S rRNA features to all the sequences, so that the "Features" page of the form will be complete when you get to it.

The "Sequence(s) and Definition Line(s)"Section

Molecule Type

You must indicate the type of molecule sequenced by picking a molecule type from the drop-down list of molecule types. You will not be able to submit your sequence(s) until you select a molecule type.

Topology

The topology of your sequence is the 3-dimensional physical form that the sequence takes as a molecule in nature. Since most of the sequences GenBank receives are linear, we have made the preset selection for this question "Linear".

When do I select "Circular"?

- Select "Circular" only when you submit **thecomplete sequence** of a circular molecule (e.g. the complete sequence of a chloroplast, plasmid or mitochondrion)
- Do not select "Circular" if you are submitting a sequence that is just a fragment or a piece of a circular molecule — in such a case you would select "Linear"

Nucleotide Sequence(s) and Definition Line(s)

In this subsection of the Nucleotide page, tell us how many sequences you are submitting, and then give us the sequences themselves.

You can give us your sequence one of two ways:

- You can paste your sequence(s) in FASTA format in the text box provided

  OR
- You can upload a file of FASTA formatted sequences from your computer directly to BankIt (Click the "Browse" button to find the file on your local computer, and click the "Upload" button to retrieve the file)

Use only one of the above methods to give us your sequence(s). Do not paste your sequence(s) into the text box **and** upload the sequence from your computer. If you do, you will get an error message and will not be allowed to submit unless you remove either the sequences you uploaded or the sequences you pasted in the text box.

Submitting Multiple Related Sequences

You can submit multiple sequences using BankIt since it is intended for the submission of simple sets. A simple set is a group of any number of sequences (70 sequences or 700 sequences) that all have the same feature or the same few features (e.g. the same CDS and the same gene).

Each member of a set must be unique. For example:

- Have unique source information (clones, isolates, strains, vouchers, etc.)
- Be a unique species (e.g. a group of 20 different spiders from the same cave)

You do not have a simple set if each of the sequences you want to submit as a group has features different from the other sequences in the group. Cases such as these are a "batch set", and require specific feature annotation later in the submission process. For example, a batch might be the sequences of 10 different genes from the same organism.

The FASTA Format

The FASTA format includes a sequence ID, source information, a single-line description of the sequence (called the definition line), and the raw sequence data:

```
>Seq3 [organism=Dendroica tigrina] myoglobin from high canopy warbler
CCTATACCTAATTTTCGGCGCATGAGCCGGAATGGTGGGTACCGCTCTAAGCCTCCTCATTCGAGCAGAA
CTAGGCCAACCCGGAGCCCTTCTGGGAGACGACCAAGTCTACAACGTGGTTGTCACGGCCCATGCCTTCG
```

The Definition Line

The FASTA definition line format is very specific so that BankIt can read the information you give in the definition line and put it in the right place within your submission. For this reason, it is important that you follow the format examples provided on the BankIt form or in the step-by-step instructions provided in the GenBank Submissions Quick Start:

After you provide the organism name in the definition line, provide as much additional text as you need to fully describe the sequence. Do NOT use definition lines that provide no information about the sequence (e.g. "Definition Line for Sequence 1").

Using Source Modifiers in the Definition Line

BankIt will read source modifiers in the definition line and will use the source information provided there to automatically fill out the source modifier section of the BankIt submission form for you if the source modifiers are formatted as follows:

[source modifier=value]

- Here are examples of source modifiers in the correct format for the definition line:

  [country=USA]
  [breed=Hampshire]
  [collected_by=T. Jones]
- Here is an example of a definition line that contains source modifiers:

```
>Seq1 [organism=Sus scrofa] [breed=Hampshire] [country=USA] [collected_by=T. Jones]
```

You will find a complete list of source modifiers in the BankIt Help documentation available online.

Uploading vs. Pasting your FASTA file:

Uploading your FASTA file

Use the "Upload FASTA file" option to submit multiple sequences. It is easier to create a FASTA file that contains many sequences and upload it than it is to cut and paste all of the sequences and then create the FASTA format for them in the text box.

Pasting your FASTA file

Use the "Paste Sequence(s)" box primarily to submit a single sequence, without creating a separate FASTA formatted file for it.

Common Mistakes Made While Filling Out the "Nucleotide" Page

- **Mistake: Entering the incorrect submission release date**

  Fix: Be sure that the release date and year are correct. You will be reminded of this date at the end of the submission process.
- **Mistake: Not using the correct format in the FASTA definition line**

  Fix: The FASTA definition line format is very specific so that BankIt can read the information you give in the definition line and put it in the right place within your submission. For this reason, it is important that you follow the format examples provided on the BankIt form or in the step-by-step instructions provided in the GenBank Submissions Quick Start:
  - **Did you put a space between your sequence identifier (ID) and the organism source modifier?**

    If you didn't, you'll get an error message that says that you have no sequence ID or that the sequence ID contains invalid characters. You won't be able to proceed with your submission until you correct the problem.
    The correct format is:
    >Seq1 [organism=Homo sapiens]
    NOT
    >Seq1[organism=Homo sapiens]
  - **Did you put a > sign in front of your sequence ID with no spaces between the sign and the Sequence ID?**

    If you didn't put the > sign in front of your sequence ID, or you put a space between the > sign and the sequence ID, you'll get an error message that says that you have no sequence ID or that the sequence ID contains invalid characters. You won't be able to proceed with your submission until you correct the problem.
  - **Did you use the correct format for the organism source?**
    The only format that BankIt will recognize for the organism source is: **[organism=value]**
    If you used any format that is different from the format shown above, like **(organism Homo sapiens)**, BankIt will not be able to read the format you used.
    - If you submitted more than one sequence:

      and made an error in the organism source format

      OR

      you did not provide the organism in the definition line in one or more of the sequences you entered in the Nucleotide page

      You will be required to either provide the organism in the definition line(s) that are missing it,

or fix the organism format error(s) in the definition line(s) before you will be allowed to go to the next page of the form.