The GenBank Submissions Handbook

Last Updated: November 3, 2014

National Center for Biotechnology Information (US)
Bethesda (MD)

National Center for Biotechnology Information (US), Bethesda (MD)

NLM Citation: The GenBank Submissions Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-.

This book contains information on GenBank, the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences.

Table of Contents

GenBank Submission Resources Quick Start

Created: April 6, 2011; Updated: November 3, 2014.

This guide will help you begin the GenBank submission process. It is arranged as a topical reference, so you do not have to read from beginning to end to find specific pieces of information. Each question/answer unit in this Quick Start will:

- Provide information that will address common GenBank submission questions
- Provide links to appropriate GenBank resources
- Provide links to specific GenBank user documentation that will address the question in greater detail.

Although most general GenBank submission questions should be addressed by the information contained in this document, if you are unable to find the information you are looking for, contact info@ncbi.nlm.nih.gov .

To begin searching this section of the GenBank Submission Resources Quick Start, you can either:

Enter your **search word(s)** text in the text box **at the top of the page** and click on the "Go" button,

OR

Click on any of the "GenBank Submission Resources Quick Start" **sub-categories listed** in the "Contents" section **below** to navigate to the sub-category of your choice.

What Kind of Data Can be Submitted to GenBank?

Created: April 6, 2011; Updated: November 3, 2014.

What kind of data will GenBank accept?

GenBank is a nucleotide sequence database and will accept primary sequence data that was directly determined by the submitter.

Below are examples of submission types included in GenBank:

- mRNA Sequences
- Prokaryotic Genes
- Eukaryotic Genes
- rRNA and/or ITS
- Viral Sequences
- Transposon or Insertion Sequences
- Microsatellite Sequences
- Pseudogenes
- Cloning Vectors
- Phylogenetic or Population Sets
- Non-coding RNAs

The following submission types are accepted by GenBank, but should be submitted using their own submission tools (see below):

- Expressed Sequence Tags (**EST**) should be submitted directly to dbEST (the EST division of GenBank)
- Genome Survey Sequences (**GSS**) should be submitted directly to dbGSS (the GSS division of GenBank)
- Transcriptome Shotgun Assembly (**TSA**) should be submitted directly through the submission portal according to these directions.

If your submission does not fall into one of the above categories, contact info@ncbi.nlm.nih.gov to determine which NCBI resource would be most appropriate for your submission.

For help beginning the submissions process to GenBank, see the "Submitting Sequences using Specific NCBI Submission Tools" section of this Quick Start.

What kind of data will GenBank NOT accept?

The following submission types are not accepted by GenBank:

- Sequences <200 bp long. Unassembled sequences from next-generation sequencing platforms should be submitted to the Sequence Read Archive (SRA)
- A genomic sequence of multiple exons joined together without the sequence of the intervening introns or without a 'gap' of internal nnns representing the missing sequence
- Primer only sequences (These sequences can be submitted directly to NCBI's Probe database)
- Protein only sequences
- Sequences containing a mix of genomic and mRNA sequence represented as a single sequence
- Sequences without a physical counterpart (consensus sequences)

For help beginning the submissions process to GenBank, see the "Submitting Sequences using Specific NCBI Submission Tools" section of this Quick Start.

Can I submit a sequence contig to GenBank?

The answer to this question depends upon the sequence contig you intend to submit.

- Sequence contigs assembled from sequence already present in International Nucleotide Sequence Database Collaboration (INSDC) sites should be submitted to NCBI's Third Party Annotation sequence database (TPA).
- Sequence contigs assembled from nucleotides that you have sequenced yourself and assembled using sequence overlap can be submitted to GenBank. If there are gaps in your contig assembly, they must be filled with internal nnns that represent any missing sequence.
- Computer-derived mRNA assemblies should be submitted to TSA.

For help beginning the submissions process to GenBank, see the "Submitting Sequences using Specific NCBI Submission Tools" section of this Quick Start.

How do I submit a large number of cosmid, BAC, or YAC derived genomic clones to GenBank?

The best way to submit a large number of cosmid, BAC, or YAC derived genomic clones to GenBank is to submit through our High-Throughput Genomic (HTG) sequence division.

The HTG division contains unfinished high-throughput DNA sequences that are available for BLAST similarity searches against the"HTGS" database.

Note: Sequences submitted via the HTG automated system are made public immediately and cannot be held for release at a later date.

If you would like more information about submitting to the HTG division of GenBank, contact the HTG division at: htgs-admin@ncbi.nlm.nih.gov .

Can I submit protein-only sequences to GenBank?

GenBank is a nucleotide sequence database and therefore **does not accept protein-only sequence submissions**.

If you do not have nucleotide sequence(s) for your protein(s), but would still like to submit your directly sequenced protein to a public database, see The Universal Protein Resource (UniProt).

Can I submit primer sequences to GenBank?

GenBank does not accept primer sequences, but you can submit primers to NCBI's Probe database, which is a public registry of nucleic acid reagents designed for use in a wide variety of biomedical research applications.

The Probe database also includes information on reagent distributors and probe effectiveness, as well as computed sequence similarities. The Probe database submission documentation provides a simple overview of the required information you will need, as well as some basic rules to follow during your submission.

Before Starting the Submission Process

Created: April 6, 2011; Updated: November 3, 2014.

Choosing the Appropriate Submission Resource

Submission Tools

BankIt vs. Sequin vs. tbl2asn

When should I use BankIt for submissions? When should I use Sequin or tbl2asn?

You should use BankIt if:

- You prefer to use a web-based submission tool
- You do not require advanced sequence analysis tools

You should use Sequin if:

- You prefer to work on your submission off-line
- You would like graphical viewing and editing options, including an alignment editor
- You would like the option to have network access to related analytical tools
- You are submitting files containing less than 10,000 sequences. If you have more than 10,000 sequences, you must submit multiple files or use tbl2asn.

You should use tbl2asn if:

- Your sequence has a lot of annotation
- You are submitting a large batch of sequences
- You have Whole Genome Shotgun (WGS) submissions or Transcriptome Shotgun Assembly (TSA) submissions
- You have complete genome submissions
- You are submitting FLIC sequences

Once you have decided which of these tools you'd like to use for your submission, see the "Submitting Sequences using Specific NCBI Submission Tools" section of this Quick Start guide for brief explanations of each of the different submission processes as well as links to useful material.

See the GenBank Sample Record page, which provides all GenBank Record field definitions.

Expressed Sequence Tags

Which submission tool should I use if I want to submit Expressed Sequence Tags (ESTs)?

ESTs should be submitted through the database of Expressed Sequence Tags (dbEST) submission system.

You may wish to look at the ""Submitting Sequences using Specific NCBI Submission Tools" section of this Quick Start guide for a brief description of the dbEST submission process, and links to useful material.

I have computationally assembled mRNA sequence reads from primary data such as ESTs, traces, and Next Generation Sequencing Techonologies. Where do I submit my assembly?

You can submit your assembly to the Transcriptome Shotgun Assembly (TSA) Sequence Database using the process described on the TSA home page.

Genome Survey Sequences

Which submission tool should I use if I want to submit Genome Survey Sequences (GSSs)?

GSSs should be submitted through the database of Genome Survey Sequences (dbGSS) submission system.

dbGSS contains (but is not limited to) genomic sequences from the following types of data:

- Random "single pass read" genome survey sequences
- Single pass reads from cosmid/BAC/YAC ends (these may or may not be chromosome specific)
- Exon trapped genomic sequences
- Alu PCR sequences

You may wish to look at the "Submitting Sequences using Specific NCBI Submission Tools" section of this Quick Start guide for a brief description of the dbGSS submission process, and for links to useful material.

Barcode Sequences

What are "Barcode" sequences, and where are they submitted?

Barcode sequences, determined as part of the Barcode of Life initiative, are short nucleotide sequences from a standard genetic locus for use in species identification. Currently, the Barcode sequence being accepted for animals is a 5', 658 base pair region of the mitochondrial cytochrome oxidase subunit I (COI) gene.

The Barcode Submission tool provides for streamlined online submission of Barcode sequences into GenBank. With this tool, you can:

- submit new Barcode sets
- complete your most recent incomplete submission
- download a flat file summary of completed submissions

First-pass sequence data generated from a single cosmid, BAC, YAC, or PAC clone

Where do I submit unfinished DNA (e.g. first-pass sequence data generated from a single cosmid, BAC, YAC, or PAC clone) sequences?

An unfinished collection of DNA sequence data derived from a single cosmid, BAC, YAC or PAC clone that may contain one or more gaps, can be submitted to the High Throughput Genomic (HTG) sequence division of GenBank. The HTG division contains unfinished high-throughput clone-based DNA sequences that are available in GenBank and for BLAST similarity searches against the "HTGS" database.

A single accession number is assigned to sequence data generated from a single clone; each HTG record provides a user with the HTG sequence status and a flag that the sequence data are "unfinished" and may contain errors.

If you want to submit data to the HTG division of GenBank, review the HTG submission documentation, as well as HTG FAQs, both of which can be accessed using links located on the HTG home page. There is a brief overview of the HTG submission process in this Quick Start as well.

Note: The HTG submission system releases all submissions immediately after processing – you cannot set a release date in advance. **If you need to set a release date for your submission, you must submit using the standard GenBank submission pathway**.

If you would like more information about submitting to the HTG division of GenBank, contact the HTG division at: htgs-admin@ncbi.nlm.nih.gov .

Flatfiles generated using a non-NCBI tool

Can I submit a flat file to GenBank created using a NCBI tool if the file generated looks just like Sequin output or looks just like a GenBank flat file?

GenBank cannot accept flat files created using non-NCBI tools for the following reasons:

- **We cannot accept the flat file format** (even if it is made in Sequin) since the flat file format is a display format only.
- **Sequin is not able to interpret GenBank-style files generated by outside tools** since GenBank requires ASN.1 formatting for proper field specification of features and other elements in a submission, which outside tools are unable to provide.

Please submit properly formatted ASN.1 files.

Learn about GenBank Records before you Submit

How do I find definitions for all the fields in a GenBank Record so I know what they are before I begin the GenBank submissions process?

To explore the definitions the fields in a GenBank Record before you start your submission, **go to our online** example **of a GenBank record**, and click on any of the light blue links to see the field definitions.

What's the difference between a GenBank record and a RefSeq record?

- **A**GenBank**record represents** primary sequence data supplied by the original submitter, who has editorial control over that record's data.
- **Reference Sequence (**RefSeq**) records** are derived from GenBank records, but differ from them in that each RefSeq is a curated synthesis of information about a particular sequence**,** rather than an archived unit of primary research data like the records in GenBank.
- The RefSeq database aims to provide a non-redundant, well-annotated set of curated sequences to be used as stable references for annotation and for various studies and analyses.
- A RefSeq record will cite the accession numbers of the original GenBank records from which it was derived.
- **RefSeq records may be altered by NCBI staff as needed** to incorporate additional sequence or annotation information. In addition, changes to an original GenBank record by its submitters may be incorporated by NCBI staff into a RefSeq record.

The complete original publication describing RefSeq is available on PubMed Central.

Submitting Sequences using Specific NCBI Submission Tools

Created: April 6, 2011; Updated: November 3, 2014.

Submission using BankIt

How do I create a submission to GenBank using BankIt?

1. **Review the**Requirements for GenBank Submissions through BankIt, and make sure you can provide the required information for your submission.
2. If you have never submitted to GenBank **scan the**GenBank Sample Record**to familiarize yourself with GenBank record field definitions**.
3. For examples of specific types of GenBank submissions, see the GenBank Annotation Example page.
4. **Login to MyNCBI:**
   a. **Go to the**BankIt home page, and click on "Sign in to use BankIt" located in a yellow box on the right at the top of the page. You will go to the MyNCBI login page.
   b. If you do not already have an NCBI account**, click on "Register for an NCBI account"** located below the login text boxes. The boxes marked with an asterisk (*) indicate the minimum amount of information we need to create an account for you. Enter the required information and click the "Create account" button.
5. **Login to BankIt and begin your submission**. The submission process has well marked steps where you will be prompted to provide contact information and your data.
6. The BankIt home page contains links to sequence annotation examples.

Submission using Sequin

How do I create a submission using Sequin?

1. **Download the Sequin program** from the Sequin Site:
   a. Go to the Sequin Home page, and click on the "Download Sequin" link located on the side-bar. You will go to the Sequin FTP site where you can download the correct file for your operating system.
   b. If you are uncertain which FTP file to use, see the "Select download type" page to get downloading instructions for your specific operating system.
   c. If you have trouble downloading or installing Sequin, see the troubleshooting guide.
2. **Prepare a properly formatted FASTA file** of your sequence data:
   a. **The FASTA format is raw sequence preceded by a definition line**:
   The definition line begins with a > sign and is followed immediately by the name of your sequence (your own local identification code, or sequence ID) and a title that describes the sequence. Be sure to use a text editor when you create your FASTA file.
   For more information on FASTA sequence formatting, see the FASTA section of the Sequin help document.
   b. Embed important information in the title portion of the definition line and Sequin will use this information to help construct your sequence record. For example:
      - **You can enter organism and strain or clone information** in the title portion of a nucleotide definition line using name=value pairs surrounded by square brackets: [organism=Drosophila melanogaster] [strain=Oregon R]
      - **You can enter gene and protein information** in the title portion of a protein definition line using name=value pairs surrounded by square brackets: [gene=eIF4E] [protein=eukaryotic initiation factor 4E-I]

3. **Launch Sequin**. During the Sequin submission process, Sequin will prompt you to provide the information we need to process your submission.
    - **Sequin has context-sensitive, on-screen help** that will open automatically when you start Sequin. Because it is context sensitive, the Help text will change and follow your steps as you progress through the program.
4. **Submit your completed Sequin file as ASN.1** rather than as a flat file:
    - Be sure to review and fix validation problems before saving your file.
    - To save the file as ASN.1, you can either click the "Done" button on the record viewer, or go to the "File" menu and select "Prepare Submission", which will also save the file as ASN.1.
    - We cannot accept the flat file format since the flat file format is a display format only.
5. When you have completed the submission process, **you must email the.sqn submission files generated by the Sequin program togb-sub@ncbi.nlm.nih.gov**, since Sequin does not automatically transmit the completed file for you at the end of the Sequin process (a dialog box will appear at the end of the Sequin submission process instructing you to email your submission files). Note: Do not encode the files before sending.
6. **When we receive a new Sequin submission**, an automatic reply will be generated and sent to the email address used to submit to GenBank. This automatic reply confirms that we received your submission, and states that you will be hearing from the GenBank submissions staff within two working days.

If I've created my submission file using Sequin, what file do I submit — a flat file or an ASN.1 file?

To save the file as ASN.1, you can either click the "Done" button on the record viewer, or go to the "File" menu and select "Prepare Submission", which will also save the file as ASN.1. Be sure to review and fix validation problems before saving the file.

We cannot accept flat file format even if it is made in Sequin since the flat file format is a display format only.

Submission using tbl2asn

When is tbl2asn a good alternative to Sequin, and can you give me step-by-step instructions for using tbl2asn to create a submission to GenBank?

tbl2asn is a program that allows a user who has:

- Large batches of sequence
- A lot of annotation
- Complete Genomes
- Whole Genome Shotgun submissions

to create a Sequin (.sqn) file for submission without having to go through the step by step process of using the Sequin program.

All you need to do to use the tbl2asn program is:

- Place your data in appropriately formatted files
- Download and run the tbl2asn program
- Command the tbl2asn program to use the data files to generate a .sqn file, which you can submit by email to GenBank.

The main difference between Sequin and tbl2asn is that Sequin is a menu driven program with a graphical user interface, while tbl2asn is a command line program where the user interacts with the tbl2asn software by typing commands to perform specific tasks.

Below are step-by step instructions for creating a .sqn file using tbl2asn. Should you need further information about any part of this process, see the tbl2asn home page or the tbl2asn documentation available on the NCBI toolbox FTP site:

A. **Place your data into appropriately formatted data files and place the data files together in a single directory:**
   1. **There are 6 types of data files that tbl2asn can use to construct a Sequin submission**. 3 are required, and 3 are optional:
      Required files:
      - Template file (see step "B" below) containing a text ASN.1 Submit-block object (use file suffix .sbt)
      - FASTA file for Nucleotide sequence data (use file suffix .fsa)
      - Feature Table file (use file suffix .tbl). [Required only if including annotation]

      Optional files:
      - Quality Score file (use file suffix .qvl)
      - Source Table file (use file suffix .src) [useful when submitting multiple records with source qualifiers that have different values]
      - Protein sequence file (use file suffix .pep) [These files are rarely needed]
   2. **The prefixes (base names) of the different files** (with the exception of the .sbt file) **you are going to use together to construct a submission should be the same as that of the .fsa file** since tbl2asn will look for .tbl, .src, and .qvl files that have the same prefix as the .fsa file in order to make the Sequin file. For example:
      - **template**.sbt (this is the only file whose prefix is different. Leave the prefix as is).
      - chr01.fsa
      - chr01.tbl
      - chr01.qvl
   3. **Save the files that you will be using in a single directoryto construct a submission in the same directory**.

B. **Create a Submission Template file**:
   (You can create this file by using the Online Submission Template page or by using Sequin.)
   1. **Using the Online Submission Template page:**
      a. Go to the online Create Submission Template page.
      b. Fill in the required (*) and optional textboxes in the "Contact Information", "Sequence Authors" and "Reference Information" sections.
      c. Click on the "Create Template" button at the bottom of the page.
      d. SAVE the file as **template.sbt** to the same directory that contains the other files to be used for the submission.
   2. **Using Sequin:**
      a. Load Sequin onto your computer.
      b. Click the "Start New Submission" button on the Sequin startup page.
      c. Enter the manuscript title if desired and click the "next page" button.
      d. Enter your contact information and click the "next page" button.
      e. Enter the author information and click the "next page" button.
      f. Enter the affiliation information and click the "next page" button.
      g. Click on the white submission tab to return to the submission window.
      h. Click on "File" located at the upper left corner of the submission window to activate a drop-down menu and select "Export Submitter Info".
      i. Save the file as template.sbt.

C. **Download the tbl2asn program appropriate to your operating system:**
   1. Go to the tbl2asn FTP site.
   2. Click on the file appropriate for your operating system to download.
   3. Uncompress the tbl2asn file using the appropriate zip utility.
   4. Rename the file **tbl2asn**, and set permissions as required for your operating system.

D. **Open a command line interpreter for Windows or Mac operating systems**. Because tbl2asn is a command line program, you can't just click on the tbl2asn icon to open it like you would a standard graphical interface program like Sequin. You first have to open a command line interpreter for your operating system, and then once you are in the command line interpreter, you can command your computer to run the tbl2asn program:
   - **In the Windows operating system (OS), the command line interface is called "Command Prompt", which you can find by doing the following**:
     a. Go to the "Start" menu.
     b. Click on "All programs" to release a menu.
     c. Click on "Accessories" to release a menu.
     d. Click on "Command Prompt" to open the "Command Prompt" command line interpreter.
   - **In the Mac operating system (OS), the command line interface is called "Terminal", which you can find by doing the following**:
     a. Go to the "Applications" folder.
     b. Double click on the "Utilities" folder to open it.
     c. Double click on "Terminal" to open the "Terminal" command line interpreter.
   - **In the Linux operating system (OS)**, open a shell and use chmod +x to allow the downloaded program to be executed. (File transfer with FTP does not retain UNIX file permissions.)

E. **Move tbl2asn.exe to a directory that your computer will search automatically when you enter the tbl2asn command**. The benefit of moving tbl2asn.exe to an automatically searched directory is that you will only have to enter the command **tbl2asn** after the prompt without having to type out a lengthy path to tbl2asn.exe every time you want to use it.

Windows OS:

a. Go to "Command Prompt" command line interpreter you opened in step D.
b. Type **path** following the prompt, and hit the "Enter" button. The computer will list all the paths (the PATH directories) it automatically searches when you enter a command in the command line interface.
c. For example:

```
C:\Documents and Settings\Owner>path
PATH=C:\WINDOWS\System32; C:\WINDOWS; WINDOWS\System32\Wbem
```

   From the above response, you can see that the for the example computer, the WINDOWS directory is one of the directories that is automatically searched when the user enters a command (the PATH directory). So, if you place tbl2asn.exe in this directory, the computer will find and run tbl2asn.exe if you type the command **tbl2asn** following the prompt.
d. Move tbl2asn to the PATH directory mentioned in previous step.

Mac OS:

a. Open the Applications folder.
b. Create a new folder, and give it a recognizable name. For this example, we'll name the folder: Command_line_apps.

c. Move the tsb2asn file you downloaded in step C into the Command_line_apps folder.
d. Go to the "Terminal" command line interpreter you opened in step D, and enter the following command:

```
export PATH=/Applications/Command_line_apps:$PATH
```

At this point, you can start the tbl2asn program in the command line interpreter by using the command **tbl2asn**.
Note: you will have to repeat step d of these instructions for each new Mac "Terminal" session in order to use the command **tbl2asn**, since the command given in step d is not remembered from one "Terminal" session to the next.

F. **Change the default directory of your command line interface to the directory that contains your data files.** The benefit of changing directories on your command line interface to the directory that contains your data files is that you can enter tbl2asn commands without having to type a lengthy path to the directory that houses the files for each command you use:
   1. Following the prompt, type cd (change directories) followed by a space, and then by the path to the directory that contains your data files. Hit the "Enter" button.
   2. The prompt should change to reflect the new directory (called Sequence_Data in the following examples):
      - **Example of the cd command followed by the new prompt** (changed to reflect the new directory) **in a Windows OS command line interpreter**. In this example, the data files are housed in a file called "Sequence_Data", which in turn are housed in a directory called "My Documents":

        ```
        C:\Documents and Settings\Owner\>cd C:\Documents and
        +Settings\Owner\My Documents\Sequence Data
        C:\Documents and Settings\Owner\My Documents
        \Sequence_Data>
        ```

      - **Example of the cd command followed by the new prompt** (changed to reflect the new directory) **in a Mac OS command line interpreter**. In this example, the data files are housed in a file called "Sequence_Data":

        ```
        Apple1:~ Username$ cd Documents/Sequence_Data
        Apple1:~/Documents/Sequence_Data Username$
        ```

        Where Apple 1 = computer name in Mac OS; this name will be different for every computer as it reflects each individual computer's name.

G. **Run tbl2asn within a command line interpreter and access all current tbl2asn commands.** A list and definitions for the most commonly used tbl2asn commands is available on the tbl2asn home page. You can access a complete list of all tbl2asn commands by doing the following (these instructions will work for all operating systems):
   1. Open the command line interpreter for your operating system.
   2. Type **tbl2asn** followed by a space and a hyphen after the command line interpreter prompt:
      Prompt>tbl2asn -
   3. Hit the "enter" key to display all the tbl2asn commands.
   4. **The following TBL2ASN commands will be used in the example below to create a sequin file** using the minimum required files: a FASTA file, a feature table, and a template file.
      - -p specifies the path for the table and sequence files [required]
      - -t specifies the template file (including the path) [required]

- -j allows the addition of source qualifiers that will be the same for each submission
  Example: -j "[organism=Saccharomyces cerevisiae] [strain=S288C]"
- -V is a verification function when combined with the following:
  v performs a validation [optional but strongly recommended]
  b generates GenBank flatfiles with a .gbf suffix
  (Sample command line: -V vb). Note: The flatfile format is for viewing only. It cannot be submitted.

H. **Use tbl2asn commands to create a Sequin (.sqn) file using required data files**. **NOTE**: If you intend to submit multiple sequences in a single submission, see Step H6 of this question before beginning. The following example will show how to use the above tbl2asn commands to create a Sequin (.sqn) file using the minimum required files —a template file (.sbt), a FASTA file (.fsa) and a feature table file (.tbl). The following instructions will work for all operating systems.

In order to use the instructions as written below, you must place tbl2asn.exe in a path directory (**see step E**), and you must change the command line interpreter's default directory to one that houses all the files you intend to use in a single submission (**see step F**).

1. **Type tbl2asn after the prompt** (do not hit "Enter" yet).
   This command tells the computer to run tbl2asn.exe
   Example:
   ```
   Prompt>tbl2asn
   ```
2. **Type a space after tbl2asn, then type –t, another space, and the name of your template file** (do not hit "Enter" yet).
   The **-t** command followed by the name of your template file tells the computer the name of the template file to use in creation of your .sqn file when using tbl2asn.
   Example:
   ```
   Prompt>tbl2asn -t template.sbt
   ```
3. **Type a space following the name of your template file, then type -p, another space, then type a period (dot)** (do not hit "Enter" yet).
   -p alone tells the computer where to look for the table and sequence files. **-p** followed by a space and then a dot tells the computer to look for the table and sequence files in the current directory.
   Example:
   ```
   Prompt>tbl2asn -t template.sbt -p .
   ```
4. **Type a space following the dot, then type –j, then type another space, and then provide source modifier information inside of quotation marks and brackets**. For example:
   "[organism=Saccharomyces cerevisiae] [strain=S288C]" (do not hit "Enter" yet)
   -j tells the computer to add the source information that follows to each submission. If there is annotation and the genetic code is not the standard code, then include the correct code in the fsa definition line, or with the -j in the command line, to avoid errors.

   Example:
   ```
   Prompt>tbl2asn -t template.sbt -p . -j
   "[organism=Saccharomyces
   cerevisiae] [strain=S288C]"
   ```

NOTE: The method stated in step 4 is good if you have source information that is common to all the files in the directory. If you have additional source information that is specific to particular submissions, omit the -j command, and:

- include the source information in the definition line of each FASTA (.fsa) sequence file.

OR

- create a tab-delimited source table (file suffix .src) for each .fsa file, and place it in the directory where the other files specific to a particular submission are housed.

5. **Type a space following the source information in quotation marks and brackets, then type–V** , another space, and then **vb** (do not hit "Enter" yet).

   -V is a verification command when used in conjunction with **v** (strongly suggested), which will tell the computer to run a validation step to insure that there are no errors in your submission.

   This validation step will generate a report (with suffix .val) for each .fsa file and place it in the same directory that houses the data files and tables used in the submission.

   If you add a **b** command (optional) following the **v** command, the computer will generate a GenBank flat file (.gbf) of your submission and deposit it in the same directory that houses the data files and tables used in the submission. Note that .gbf files are not suitable for submission. They are only to view the file in GenBank flatfile format.

   Example:

```
Prompt>tbl2asn -t template.sbt -p . -j
"[organism=Saccharomyces
cerevisiae] [strain=S288C]" -V vb
```

6. **Optional Step**
   If your submission contains multiple records, put them in a single .fsa file, and add the following to the command line:

   Type a space following your last command, then type–athen type another space and then type s
   Example:

```
Prompt>tbl2asn -t template.sbt -p . -j "[organism=Saccharomyces
cerevisiae] [strain=S288C]" -V vb -a s
```

The –a command used in conjunction with the s command instructs tbl2asn to read multiple FASTA components in one file as a set of unrelated sequences. This creates a single file of multiple submissions.

Remember: Each .tbl and .qvl file will need to contain the information for the sequences in the corresponding .fsa file.

Note: You will have to try to achieve a balance between the number of files you submit and the number of records per file submitted. In general, limit the number of records (sequences) per file to a few thousand when there is no annotation, and a few hundred when there is annotation.

7. **Hit the "Enter" key on your keyboard**.
   The response will be as follows:

   [**tbl2asn** and the current tbl2asn version number] **Flatfile** followed by the prefix used for the tables and files used for this specific submission.

   Followed by

   [**tbl2asn** and the current tbl2asn version number] **Validating** followed by the prefix used for the tables and files used for this specific submission.]

   Followed by

   the command line interface prompt.
   Example:

```
[tbl2asn 15.2] Flatfile chr01
[tbl2asn 15.2] Validating chr01
Prompt>
```

8. **Find the Sequin file (.sqn), the validation file (.val) and the GenBank Flatfile (.gbf) generated by the tbl2asn program for each .fsa file**:

   Once tbl2asn generates the .sqn, .val and .gbf files, it automatically places them in the same directory that houses the data files and tables used in the submission.

   I. **Open the .val file using a text editor to see the errors. Open the .sqn file in Sequin to correct any errors that are mentioned.** Or you can make the appropriate changes in the .tbl file and remake the .sqn file.
      - Taxonomy-related errors about missing lineages can be ignored
      - If there is annotation and the genetic code is not the standard code, then include the correct code in the .fsa definition line, or with the -j in the command line, to avoid errors

   J. **Optional:** open the Genbank Flatfile with a text editor (if you chose to include the generation of a GenBank Flatfile in your commands) and review. The .gbf files are only for display, not for submission. Do not make any changes to these files; the changes need to be in the .sqn files.

   K. **After you have checked your files and corrected any errors that the validation step found, send the files to Genbank** either by email to gb-sub@ncbi.nlm.nih.gov, or by using SequinMacroSend for regular submissions, and GenomesMacroSend for genome (complete or incomplete whole genome shotgun) submissions.

When submitting to GenBank using tbl2asn, do I always have to provide a project ID?

The answer to this question depends upon what you are submitting.

A project ID is only required if you are submitting:

- A bacterial or eukaryotic (non-organelle) genome
- Genome-scale studies including:
  - Targeted loci studies
  - Metagenomic studies
  - Multi-isolate studies

For information on how to register your project with BioProject and get a Project ID, see the "Submitting to BioProject" section of the BioProject help manual.

Submitting Different Sequence Types using Specific NCBI Submission Resources

Created: April 6, 2011; Updated: November 3, 2014.

Submission to the Transcriptome Shotgun Assembly Sequence database (TSA)

Can I submit a Transcriptome Shotgun Assembly (TSA) that I have assembled out of sequences I found in dbEST, the Sequence Read Archive, or the Trace archive?

No, you can submit a TSA sequence only if you have experimentally determined the primary sequences used to assemble the TSA sequence yourself.

How do I submit Transcriptome Shotgun Assembly (TSA) Sequences?

Complete details of the TSA submission process are available on the TSA home page.

Submission to the High-Throughput Genomic (HTG) sequence division of GenBank

How do I submit to GenBank via the High-Throughput Genomic (HTG) sequence division?

Complete discussion of the HTG division of GenBank is available on the HTG home page. HTG submission instructions are available on the HTG submission page.

The HTG submission system is designed for high-throughput bulk submissions, typically of BAC clone sequences from genome centers. An automated system processes the submissions daily and releases them to GenBank the same day if there are no errors.

- **If you are only sequencing a few BAC clones do not use the HTG system** – submit using the standard GenBank submission pathway.
- The HTG submission system releases all submissions immediately after processing – you cannot set a release date in advance. **If you need to set a release date for your submission, you must submit using the standardGenBank submission pathway**.

If you have questions or are unsure of how to proceed, we recommend you contact htgs-admin@ncbi.nlm.nih.gov for advice before you begin your submission.

Below is a brief overview of what to expect in the HTG submission process:

Before you begin the HTG submission process, review the HTGS Getting Started page to see assumptions that NCBI has for HTG submitters and their data.

1. **Sequencing center or group must contacthtgs-admin@ncbi.nlm.nih.govand request an FTP account**, where you will transfer your sequence records once they are prepared for submission.
2. **Prepare your sequence for HTG submission**:
    a. An HTG submission must be in ASN.1 format.
    b. There are two NCBI tools available for creating an ASN.1 formatted HTG submission:
        i. Sequin
        Sequin contains a setting that allows genome centers to prepare HTG submissions. The user will import a FASTA formatted sequence file and a Sequin submission template file (containing contact and citation information) into Sequin, and will then enter annotation for the sequence into the Sequin form. Sequin will then generate the ASN.1

file for HTG submission.
Please see the "Formatting Sequences in FASTA Format" section of the HTG tbl2asn instructions for more instructions on creating a modified FASTA file that will define segments and gaps.
Please see the "Using Sequin to prepare a HTG Submission" page for detailed "how to" information.

ii. tbl2asn
Tbl2asn is a command line program that is downloaded to the user's computer. Once tbl2asn is installed, the user runs the tbl2asn software by typing commands that will generate an ASN.1 file for submission from a FASTA sequence file and a Sequin submission template file (containing contact and citation information).

Please see the "Formatting Sequences in FASTA Format" section of the HTG tbl2asn instructions for more instructions on creating a modified FASTA file that will define segments and gaps.

The advantage of using tbl2asn over Sequin for creating HTGS submissions is that it can be set up by the user to create submissions in bulk from multiple files.

Please see the tbl2asn section of this Quick Start for step-by-step instructions for using tbl2asn. You can find specific instructions for the HTG submission arguments required for HTG submissions on the HTG page .

c. Make sure that you include the genome center name and the sequence name in your submission; NCBI will issue the accession number. These three identifiers are required for every update, as described on the HTGS Getting Started page.

3. **Submit your HTG sequence submission ASN.1 files to NCBI into the SEQSUBMIT directory of the FTP account you requested in step 1**.
4. **Note**: HTG sequences submitted to the FTP site are processed on a daily basis.

Submission of Full Length Insert cDNA (FLIC) Submissions

How do I submit Full Length Insert cDNA (FLIC) sequence data to GenBank?

FLICs are processed via an automated FLIC processing system. **Follow the steps below to submit FLIC(s) to GenBank:**

1. To submit sequences in bulk to the FLIC processing system, a center or group must **contactgb-admin@ncbi.nlm.nih.govand request an FTP account for FLIC submissions**.
2. **Use thetbl2asnprogram to generate your submission(s)**
Submissions to the FLIC processing system must contain the following identifiers:
   - The genome center tag
   (assigned by NCBI and is generally the FTP account login name)
   - The sequence name (SeqID)
   (a unique identifier that is assigned by the submitter to a particular clone or entry and must be unique within the group's FLIC submissions)
3. **Deposit your submission(s) in the FLICSEQSUBMIT directory of your FTP account** and contact gb-admin@ncbi.nlm.nih.gov to let us know your submission(s) is available for processing.
4. **Your submission(s) will be processed and assigned an accession number**. The files will be automatically loaded into GenBank if there are no errors in the submission. FLIC submissions cannot be held confidential.

5. **Should your submission fail FLIC processing**, you will receive an email from us that describes the problem. Once you have reviewed our email, you will need to submit a corrected entry.
6. At the completion of processing, a submission report is automatically generated and deposited in your FTP account.
7. All updates to your FLIC submission(s) must include the center tag, sequence name, and accession number, or processing will fail.

Genome Submissions

Which GenBank resource do I use to submit an incomplete genome that was sequenced using clone-based sequencing, and which resource do I use to submit one that was sequenced using the Whole Genome Shotgun (WGS) approach?

Genome submissions are comprised of genomic DNA sequences representing either complete or incomplete genomes from both prokaryotes and eukaryotes. Viral, phage or complete organellar sequences should be submitted as regular GenBank records. In addition, data consisting of only a subset of genes (eg 16S ribosomal RNA or "your gene of interest"), should be submitted as regular GenBank submissions. Unassembled sequences from next-generation sequencing platforms should be submitted to the Sequence Read Archive (SRA).

WGS genomes should be submitted to the Whole Genome Shotgun (WGS) division. The clones (usually BACs) of genomes sequenced using traditional clone-based sequencing should be submitted to:

- The High-Throughput Genome (HTG) division of GenBank if they do not need to be confidential

OR

- GenBank if they do need to be kept confidential.

Whole Genome Shotgun Submissions

How do I submit a genome to GenBank via the Whole Genome Shotgun (WGS) division?

Complete submission details for the WGS division of GenBank are available on the WGS submission page. If you are unsure about the type of data submitted to the WGS division, visit the WGS List for example projects. You must register your project with the **BioProject** and **BioSample** databases**.** For further details on how to create a WGS submission, see the "How to Submit WGS Genomes" page of the WGS site".

Complete Genome Submissions

How do I submit a complete genome to GenBank?

Details details for complete, prokaryotic genomes are available in the Prokaryotic Annotation Guide. You must register your project with the BioProject and BioSample databases.

Metagenome Submissions

How do I submit a Metagenome to GenBank?

Metagenomics is the culture-independent genomic analysis of a community of microorganisms. Complete details for submitting Metagenomes to GenBank are available from the Metagenomes Submission Guide. You must register your project with the BioProject and BioSample databases.

Submission of Third Party Annotation Records

What is the Third Party Annotation (TPA) database, and what kind of data can be submitted to it?

The Third Party Annotation (TPA) database contains nucleotide sequences that are derived or assembled from primary sequence data that is housed in other International Nucleotide Sequence Collaboration (INSDC) databases (TPA is part of INSDC).

- **The INSDC databases (DDBJ,EMBL, andGenBank) containprimary sequence data** and corresponding annotations submitted by the laboratories that completed the original sequencing.
- **The TPA subset database contains nucleotide sequences built from the existing INSDC primary sequence data** that has new feature annotation described in a peer-reviewed scientific journal.

Each TPA record can be one of two types:

- Experimental: annotation is supported by wet-lab evidence published in a peer-reviewed scientific journal
- Inferential: annotation is inferred from other work by the submitter, but is not the subject of direct experimentation itself. Supporting information for the inferred annotation has been published in peer-reviewed scientific journal(s).

For additional information about what constitutes an Experimental or an Inferential TPA record, see the TPA FAQ.

A brief **list of TPA sequence submission examples can be found on** the TPA home page (scroll down a little to see the list) **as can a list of sequences/annotation that should not be submitted to the TPA** (scroll to bottom of page).

How do I submit annotation to TPA?

You can submit sequence and new annotation to the TPA database using either BankIt or Sequin, but be sure to adhere to the following guidelines regardless of which method you use to submit to TPA:

- The entire sequence submitted to TPA must be derived or built from primary sequence data.
- There is no limit on the number of overlapping/adjoining primary sequences that can be cited for a TPA submission.
- If sections of a sequence submitted to TPA have been newly determined by the submitter, those sections of sequence (if they are more than 50 nucleotides) must first be submitted to GenBank, processed, and released to the public before they can be cited as primary sequences for TPA.
- Each TPA sequence must cite the same organism as the primary sequence data used to build or derive it.

To submit to TPA using BankIt:

1. Fill out or verify your contact information on the “Contact” page, and then use the “Reference” and “Nucleotide” pages enter your publication information and your Nucleotide data.
2. On the Submission Category page, choose Third Party Annotation.
3. Follow the instructions and enter:
    a. A brief explanation of the work done as evidence to support the new feature annotation.
    b. The GenBank/INSDC accession numbers of all primary data used to assemble or derive the sequence.
4. Be sure to add all new feature annotation on the Features pages.
5. Confirm your submission on the “Review and Correct” page and click the “Finish” button to submit your sequence and annotation.

6. The submission will be labeled as a TPA record and processed accordingly after it is successfully submitted.

To submit to TPA using Sequin:

1. Begin the Sequin submission process as directed, and complete the "Submitting Authors" form. When you complete the final (Affiliation) section of the "Submitting Authors" form, click the "Next Form" button at the bottom of the page. You will go to the "Sequence Format" form.
2. Select the radio button that marks "Third Party Annotation" as your selection from the submission category section in the Sequence Format window.
3. Provide a brief explanation of the work done as evidence to support the new feature annotation in the TPA Evidence box.
4. Enter the GenBank/INSDC accession numbers of all primary sequences used to assemble or derive the TPA sequences into the "Assembly Tracking" box that appears with the flatfile display following the Annotation page.
5. Click on Accept; a new COMMENT field will appear in the flatfile, which will list the primary sequence Accession Numbers.
6. Complete your submission as directed by Sequin.
7. When you have completed the submission process, you must email the.sqn submission files generated by the Sequin program to gb-sub@ncbi.nlm.nih.gov since Sequin does not automatically transmit the completed file for you at the end of the Sequin process (a dialog box will appear at the end of the Sequin submission process instructing you to email your submission files).
8. Include a note in the email that contains the .sqn file that the submission is intended for TPA.

Submission to the database of Expressed Sequence Tags (dbEST)

What are ESTs and where do I submit them?

GenBank defines Expressed Sequence Tags (ESTs) as short (300-500 bp) single reads from mRNA (cDNA) that are usually produced in large numbers. They represent a snapshot of what is expressed in a given tissue, and/or at a given developmental stage. They also represent tags (some coding, others not) of expression for a given cDNA library.

You can submit ESTs to dbEST.

In addition to short sequence reads, **dbEST also includes** sequences that are longer than the traditional ESTs or are produced as single sequences or in small batches — including **products of differential display experiments and RACE experiments**.

dbEST is reserved for single-pass reads. **Assembled sequences should not be submitted through dbEST**, but through the TSA (Transcriptome Shotgun Assembly) division of GenBank. See the "Starting a Submission to TSA sequence database" section of this Quick Start for more information on submitting to TSA.

How do I submit to dbEST?

1. All EST sequences must be submitted using the custom streamlined submission procedures as described on the dbEST submission page. The file format used to submit to dbEST is outlined on the file format page, where you will find detailed examples of this format in use following each file format description.
2. Information (if applicable) you will be required to provide for your EST submissions will include:
   - Clone name
   - Clone library [catalog number, reference, lab source, and/or specific (in-house) name or number]

    - Tissue type
    - Developmental stage
3. Generally speaking, no annotation is expected in an EST record regardless of length, quality, or quantity of sequence submitted.
4. Once you have completed your submission files as described in the dbEST submission documentation, you will send them to: batch-sub@ncbi.nlm.nih.gov either attached to a single email message, or you can include the files in the body of the email message. Be sure that the files are in plain text (ASCII) format.

dbEST offers a specified release date option for your data should you require it. See the "Assignment of GenBank Accession Numbers and Release of Data" section of the dbEST submission page for more details.

Submission to the database of Genome Survey Sequences (dbGSS)

What are GSSs and where do I submit them?

dbGSS contains (but is not limited to) genomic sequences produced from the following:

- Random "single pass read" genome survey sequence experiments.
- Single pass reads of cosmid/BAC/YAC ends (these may or may not be chromosome specific)
- Exon trapping experiments
- Alu PCR experiments

How do I submit Genome Survey Sequences?

1. **All GSS sequences must be submitted using the custom streamlined submission procedures as described on thedbGSS submission page**. The file format used to submit to dbGSS is outlined on the file format page, where you will find detailed examples of this format in use following each file format description.
2. **Generally speaking, no annotation is expected in a GSS record** regardless of length, quality, or quantity of sequence submitted.
3. Once you have completed your submission files as described in the dbGSS submission documentation, send them to: batch-sub@ncbi.nlm.nih.gov either attached to a single email message, or you can include the files in the body of the email message. Be sure that the files are in plain text (ASCII) format.

dbGSS offers a specified release date option for your data should you require it. See the "Assignment of GenBank Accession Numbers and Release of Data" section of the dbGSS submission page for more details.

Submitting Multiple Sequences as a Set

Created: April 6, 2011; Updated: November 3, 2014.

What is the difference between a batch and a set?

A Set is a group of sequences of the same gene and/or locus that are part of a population, mutation, phylogenetic or environmental study, and are published together in the same journal article. The sequences in a Set will be processed together as a unit and will be accessible together using Entrez PopSet.

A Batch is a group of sequences that are related in some way, but from different genes and/or loci, such as coming from the ame organism or being published together in the same journal article.

When can multiple sequences be submitted together as a set?

- Members of a sequence set are published together in the same journal article, and are used to analyze the relatedness of a (sequence) population by analysis of the same gene or locus.
- Below is a list of standard set types submitted to GenBank:
  - Population study:
    a set of sequences that were derived by sequencing the same gene from different isolates of the same organism.
  - Phylogenetic study:
    a set of sequences that were derived by sequencing the same gene from different organisms.
  - Mutation study:
    a set of sequences that were derived by sequencing multiple mutations of a single gene.
  - Environmental samples:
    a set of sequences that were derived by sequencing the same gene from a population of unclassified or unknown organisms.

In all the cases listed above, each sequence set is handled as a single submission, but each sequence submitted in a set will receive its own Accession number, and can be annotated independently.

Sequence sets can be submitted to GenBank via Sequin or BankIt.

How do I submit multiple sequences as a set?

To submit a group of sequences as a set, do the following:

- **Sequin**:
  1. Fill out the Sequin Submitting Authors form.
  2. Select the set type on the Sequence Format Form (see the answer to "When can multiple sequences be submitted together as a set?).
  3. Sequin will prompt you to enter the data appropriate to the set.
- **BankIt**:

  Once you have completed steps 1 (Contact) and 2 (Reference) of the BankIt submission process, step 3 (Submission Type) will give you a choice as to the submission type you intend to submit:
  1. If you intend to submit a set, select the radio button in front of "Set".
  2. A question asking the number of sets you intend to submit will appear as will a list of set types (population/phylogenetic/mutation/environmental).
  3. Enter the total number of sets you intend to submit. (Note: you can only send one set per BankIt submission, so you will have to open a BankIt submission session for each additional set you intend to submit.).
  4. Select the radio button in front of the set type you intend to submit.

5. From this point, BankIt will ask you for appropriate source and feature data for the set type you wish to submit.

Adding Value to your Submission

Created: April 6, 2011; Updated: November 3, 2014.

Information to include with Genomic Sequence Submissions

Information to include with Eukaryotic Protein Coding Gene Sequence Submissions

I have eukaryotic genomic sequence(s) that I'd like to submit; is there additional information that I should include with the sequence(s) in my submission?

Provide the following information with your eukaryotic sequence submission:

- CDS feature(s) with product name(s), nucleotide locations, and amino acid translations of all coding regions (showing start and stop codons, if present, and the locations of any exons)
- Gene symbol(s), if known

If any of this information is not known, inform us at the time of your submission

See an online example of eukaryote genomic sequence submission annotation or perform a BLAST search to find examples of similar sequences with complete annotation

Information to include with Prokaryotic Gene Sequence Submissions

I have bacterial/archaeal genomic sequence data that I'd like to submit; is there additional information that I should include with the sequence(s) in my submission?

Provide the following information with your bacterial/archaeal sequence submission:

- **Tell us if the bacterium/archaeon is culturedoruncultured**:
  - pure culture: a culture that contains only one microbial species. Include a strain identifier.
  - enrichment culture: use of selective culture media to enrich for a set of microorganisms with a particular phenotypic property, resulting in a partially purified, mixed culture of more than one microbial species
  - uncultured bacteria/archaea are PCR-amplified directly from source/host DNA using universal primers or species-specific primers. Include the isolation_source (environmental conditions) and a unique clone identifier for each sequence.

Information to include with Prokaryotic Protein Coding Gene Sequence Submissions

I have bacterial/archaeal genomic sequence data that I'd like to submit; is there additional information that I should include with the sequence(s) in my submission?

Provide the following information with your bacterial/archaeal sequence submission:

- CDS feature (s) with product name(s), nucleotide locations, and amino acid translations(s) of all coding regions (showing start and stop codons, if present)
- Gene symbol(s), if known

If any of this information is not known, inform us at the time of your submission.

See an online example of bacterial/archaeal genomic sequence submission annotation.

Information to include with Viral Sequence Submissions

I have viral sequence data that I'd like to submit; is there additional information that I should include with the sequence(s) in my submission?

Provide the following information with your viral sequence submission:

- A unique name to distinguish each sequence you submit, such as one of the following:
  - Strain
  - Isolate
  - Clone
  - Other unique name
- Country where virus was collected (if known)
- Host (scientific/binomial or common name, if known)
- Collection date (if known)
  (use three letter abbreviation for month and four digit format for year, e.g. Feb-2001)
- Serotype or genotype (if known)
- CDS feature(s) with product name(s), nucleotide locations, and amino acid translation(s) of all coding regions (showing start and stop codons, if present)
- Gene symbol(s), if known

The information listed above should be applied to any virus submission.

If no coding region is present, provide another description of the sequence

If any of this information is not known, inform us at the time of your submission.

See an online example of viral sequence submission annotation.

Information to include with Genomic Sequence containing Structural RNA and/or Spacers

I have genomic sequence that contains structural RNA and/or spacers that I'd like to submit; is there additional information that I should include with the sequence(s) in my submission?

Provide the following information with the sequence that contains structural RNA and/or spacers:

- The names of any structural RNAs (e.g. tRNA-Ile, 16S ribosomal RNA) present
- The names of any spacer regions (e.g. internal transcribed spacer 1, 16S/23S intergenic spacer)
- The nucleotide spans of each of the above features (if known)

If you do not know the exact nucleotide spans of the above features on your sequence, tell us which of the above components you think exists in each of your sequences using a "misc_feature" or "misc_RNA" feature with the components present listed in a note.

See an online example of structural RNA and/or spacer annotation.

Information to include with promoter/genomic 5' flanking sequence/ genomic 3' flanking Sequence Submissions

I have promoter/genomic 5' flanking sequence/genomic 3' flanking sequence data that I'd like to submit; is there additional information that I should include with the sequence(s) in my submission?

Provide the following information with your promoter/**genomic 5' flanking sequence/genomic 3' flanking sequence** submission:

- The protein and/or gene symbol for the sequence to which the promoter or flanking region belongs
- Intervals of any transcribed regions (ie noncoding and/or coding mRNA exons) or coding regions, if present

If any of this information is not known, inform us at the time of your submission.

See an online example of promoter/**genomic 5' flanking sequence/genomic 3' flanking sequence** submission annotation.

Information to include with Cloning Vector Sequence Submissions

I have cloning vector sequence data that I'd like to submit; is there additional information that I should include with the sequence(s) in my submission?

Provide the following information with your cloning vector sequence submission:

- Unique name for the vector and type (ie cloning vector, expression vector, shuttle vector, etc.)
- Coding region intervals (if known) including start and stop codons
- Protein names (if known)
- Gene symbols (if known)
- Miscellaneous feature with descriptive note for biologically important regions (multiple cloning site, tags, enhancers, fusions, etc.)

If any of this information is not known, inform us at the time of your submission.

See an online example of cloning vector **sequence** submission annotation.

Information to include with Transposon/Insertion Sequence Submissions

I have Transposon/Insertion sequence(s) that I'd like to submit; is there additional information that I should include with the sequence(s) in my submission?

Include the following information with your Transposon/Insertion Sequence submission:

- The name of the transposon/insertion sequence
- The nucleotide spans corresponding to the transposon/insertion sequence
- The name of any host gene/product disrupted by the transposon/insertion sequence (if known)
- The name and nucleotide intervals of any gene/product in the transposon/insertion sequence, such as transposase (if known)
- The nucleotide spans any additional features present, such as LTRs or repeat regions (if known)

If any of this information is not known, inform us at the time of your submission.

See an online example of Transposon/Insertion sequence submission annotation.

Information to include with Microsatellite Sequence Submissions

I have microsatellite sequence data that I'd like to submit; is there additional information that I should include with the sequence(s) in my submission?

Provide the following information with your microsatellite sequence submission:

- A unique microsatellite/clone name for each sequence
- The interval of any repeat region(s) within the microsatellite sequence (if known)

Please make sure to remove all cloning vector contamination as failure to do this will delay the processing of your submission.

If any of this information is not known, inform us at the time of your submission.

See an online example of microsatellite sequence submission annotation.

Information to include with Sequence Submissions containing Repeat Regions

I'd like to submit sequence data that contains repeat regions; is there additional information that I should include with the sequence(s) in my submission?

P**rovide the following information** with your sequence submission:

- Repeat region intervals
- Repeat family, if known (eg, Alu, Mer)
- Repeat type (tandem, inverted, flanking, terminal, direct, dispersed, or other)
- Repeat unit description/intervals, if region contains more than one repeat

If any of this information is not known, inform us at the time of your submission.

See an online example repeat region sequence submission annotation.

Information to include with Pseudogene Sequence Submissions

I'd like to submit pseudogene sequence data; is there additional information that I should include with the sequence(s) in my submission?

P**rovide the following information** with your pseudogene sequence submission:

- gene intervals
- gene symbol
- protein name as a note on the gene feature

If this information is not known, inform us at the time of your submission.

See an online example of pseudogene sequence submission annotation.

Information to include with RNA/mRNA Sequence Submissions

Information to include with non-coding RNA

I have non-coding RNA sequence data that I'd like to submit; is there additional information that I should include with the sequence(s) in my submission?

Provide the following information with your non-coding RNA (ncRNA) sequence submission:

- ncRNA intervals
- ncRNA class
  You will be able to select a class from a list of options provided in both BankIt and Sequin
- ncRNA product name

If any of the above information is not known, inform us at the time of your submission.

Information to include with mRNA Sequence Submissions

I have mRNA sequence data that I'd like to submit; is there additional information that I should include with the sequence(s) in my submission?

Provide the following information with your mRNA sequence submission:

- Coding region intervals including start and stop codons, if present
- Protein name
- Gene symbol, if known
- Amino acid sequence

If any of this information is not known, inform us at the time of your submission.

See an online example of mRNA **sequence** submission annotation.

Information to Include with Alternative mRNA Transcript Submissions

I have alternative mRNA transcripts that I'd like to submit; is there additional information that I should include with the sequences in my submission?

- **Each alternative mRNA sequence must be submitted separately**. Each will be processed as a separate submission, and assigned a unique accession number.
- **Provide each alternative mRNA sequence with a specific name**. You can do this by using the name for the transcript's CDS product. For example:
  /product=actin isoform A
  /product=actin isoform B
- **Provide each alternative mRNA sequence with a note on the CDS feature** identifying the sequence as alternatively spliced:
  /note=alternatively spliced

Formatting your Submission

Created: April 6, 2011; Updated: November 3, 2014.

Sequence Formatting

FASTA Formatting

What is FASTA formatting

Define the FASTA format.

Simply speaking, the FASTA format consists of a single-line description of the sequence (called the definition line), which is followed by raw sequence data.

FASTA Formatting For Nucleotide Sequences

How do I format my nucleotide sequence in FASTA?

1. Open a text editor:
   a. Microsoft (MS) operating system:
      Open Notepad (Wordpad) or other MS compatible text editor.
      - Save document as plain text (.txt).
      - Do not save as rich text (.rtf) or as a document (.doc)
   b. Mac operating system:
      Open Textedit, TextWrangler or other Mac compatible text editor.
      - Save document as plain text (.txt).
      - Do not save as rich text (.rtf) or as a document (.doc).
2. Create the definition line for your sequence.
3. Press the “Enter” key to begin a new line following the definition line.
4. Enter your raw sequence data:
   a. Total sequence size must be at least 200 bp (shorter sequences will not be processed).
   b. Present your sequence data using Making Tab-delimited Tables symbols. Any non-IUPAC symbols (including dashes) will be removed from the sequence when it is imported into Sequin. BankIt will not accept sequences with non-IUPAC sumbols
   c. Letter case is ignored, so you may enter the sequence symbols in either upper or lower case. Any significance you attach to the case of the symbols will be lost.
   d. Use the IUPAC approved "N" symbol to represent ambiguous sequence data.
   e. Lines of sequence data in FASTA format should be 80 characters or shorter.

How do I format a FASTA definition line for a nucleotide sequence?

The FASTA definition line must be constructed in the following order and in a single line of text. Do not insert any hard returns (the “Enter” key on your keyboard) in the definition line. In both Sequin and BankIt, the organism and any additional modifiers can be added as tables later in the submission process if you do not include them in the FASTA definition line. The SeqID is the only mandatory component of the FASTA definition line.

1. Type a carat (">") symbol.
2. Enter the sequence identifier (SeqID). This SeqID:
   - Must be unique for each nucleotide sequence
   - Cannot contain any spaces
   - Cannot contain brackets

- Should be relatively short (preferably under 25 characters). Do not use the complete organism name for the SeqID.
- Isolate, strain, clone or other laboratory identifiers are examples of SeqIDs.

3. Type a space.
4. Type information about the organism where you obtained the sequence:
   - This information must be in the format [modifier=text]. For example:

     [organism=Gallus gallus]
   - Do not put spaces around the "="
5. At this point you can add additional information to describe the sequence in the form of optional modifiers:
   a. Type a space following the information you entered in step 4.
   b. Enter optional modifiers to describe the sequence:
      - This information must be in the format [modifier=text]. For example: [breed=booted bantam]
      - Do not put spaces around the "="
6. At this point, you can add an optional title for your sequence:
   a. Type a space.
   b. Enter an optional descriptive title for your sequence:
      - Here is an example of a descriptive title for a sequence: Gallus gallus doublesex and mab-3 related transcription factor 1 (DMRT1)
      - As GenBank has a preferred format for nucleotide and protein titles, the sequence title you provide will be changed to the proper format by the database staff during processing.
7. End your definition line by pressing the "Enter" key on your keyboard to insert a hard return.
8. Here is an example of a completed FASTA nucleotide sequence definition line whose components are the examples used in the steps above:
   ```
   >SEQ1 [organism=Gallus gallus] [breed=booted bantam] doublesex and
   mab-3 related transcription factor 1 (DMRT1)
   ```
9. Begin entering your raw sequence data in the accepted FASTA format

Here is an example of how the sequence definition line will look followed by nucleotide sequence:

```
>SEQ1 [organism=Gallus gallus] [breed=booted bantam] doublesex and mab-3 related
transcription factor 1 (DMRT1)
CCGGCGGCGGGCAAGAAGCTGCCGCGTCTGCCCAAGTGTGCCCGCTGCCGCAACCACGGCTACTCCTCGC
CGCTGAAGGGGCACAAGCGGTTCTGCATGTGGCGGGACTGCCAGTGCAAGAAGTGCAGCCTGATCGCCGA…
```

Remember:

- Do not include any hard returns in your FASTA definition line (by hitting the "Enter" button on your keyboard) until the end of the definition line, or you may have trouble importing your FASTA sequences to GenBank.
- If you do have trouble importing your sequences, please double check that no returns were added to the FASTA definition line by your editing software.

What kind of information should I include in the definition line of my submission?

Although you can include any information you want in your definition line**,** the information you do include will be edited during GenBank processing to conform to specific database criteria, and therefore the definition line you provide will probably not remain the same after processing**.**

See the Sequin help documentation sectionhttp://www.ncbi.nlm.nih.gov/Sequin/QuickGuide/sequin.htm - NucleotidePage on importing nucleotide FASTA for more specifics on formatting a FASTA sequence importation file.

FASTA Formatting For Protein Sequences

How do I format my protein sequence in FASTA?

1. Open a text editor:
   a. Microsoft (MS) operating system:
   Open Notepad (Wordpad) or other MS compatible text editor.
      - Save document as plain text (.txt).
      - Do not save as rich text (.rtf) or as a document (.doc).
   b. Mac operating system:
   Open Textedit, TextWrangler or other Mac compatible text editor.
      - Save document as plain text (.txt).
      - Do not save as rich text (.rtf) or as a document (.doc).
2. Create the definition line for your sequence.
3. Press the "Enter" key to begin a new line following the definition line.
4. Enter your raw sequence data:
   a. Present your sequence data using IUPAC symbols.
   b. Letter case is ignored, so you may enter the sequence symbols in either upper or lower case. Any significance you attach to the case of the symbols will be lost.
   c. Lines of sequence data in FASTA format should be 80 characters or shorter.

How do I format a FASTA definition line for a protein sequence?

Note: You will not need to import a protein sequence into BankIt or Sequin if the nucleotide spans for the coding region are provided.

The FASTA definition line must be constructed in the following order and in a single line of text. Do not insert any hard returns (the "Enter" key on your keyboard) in the definition line.

1. Type a carat (">")symbol, then a space.
2. Enter the sequence identifier (SeqID). This SeqID:
   - must be the same SeqID that you used to identify the nucleotide sequence.
   - In the case of alternatively spliced genes, a single protein FASTA file can contain two unique sequences that have the same SeqID. Both coding regions will be added to the same nucleotide sequence.
   - Cannot contain any spaces
   - Cannot contain brackets
3. Type a space.
4. Enter the protein name in the format [modifier=text]. For example:
   `[protein=doublesex and mab-3 related transcription factor 1]`
   - Do not put spaces around the "="
5. At this point you can add additional information to describe the sequence in the form of optional modifiers:
   a. Type a space following the information you entered in step 4.
   b. Enter optional modifiers to describe the sequence:
   The modifiers available for use in a protein FASTA definition line are different than those for a nucleotide FASTA definition line and are limited to the following information about the protein or gene itself, and should be presented in the format [modifier=text]:

        - [gene=text] Example:
          `[gene=DMRT1]`
        - Do not put spaces around the "="
6. End your definition line by pressing the "Enter" key on your keyboard to insert a hard return.
    - Here is an example of a completed FASTA protein sequence definition line whose components are examples used in the steps above:
      >SEQ1 [gene=DMTR1] [protein=doublesex and mab-3 related transcription factor 1]
7. Begin entering your raw sequence data in the accepted FASTA format.
   Here is an example of how the sequence definition line will look followed by protein sequence:

```
>SEQ1 [gene=DMTR1] [protein=doublesex and mab-3 related
transcription factor 1]
PAAGKKLPRLPKCARCRNHGYSSPLKGHKRFCMWRDCQCKKCSLIAERQRVMAVQVALRRQQAQEEEL
GI
SHPVPLPSAPEPVVKKSSSSSSCLLQDSSSPAHSTSTVAAAAASAPPEGRMLIQDIPSIPSRGHLEST
SD...
```

Remember:

- Do not include any hard returns in your FASTA definition line (by hitting the "Enter" button on your keyboard) until you reach the end of your definition line, or you may have trouble importing your FASTA sequences to GenBank.
- If you do have trouble importing your sequences, please double check that no returns were added to the FASTA definition line by your editing software.

Formatting Sequence Gaps

How does GenBank define a sequence "gap"?

GenBank defines a sequence gap as

- A region of unknown sequence.

OR

- A region of un-sequenceable sequence that lies between two known regions of sequence.

If I don't know the base at a particular position in my sequence data, can I use "-" or "?" to represent the unknown base?

You may use the - or ? characters in sequence data for alignment submissions only.

These symbols will be stripped from your sequence by our submission processing software if you include them in a FASTA file, so **you'll need to insert a series of nnnns where each** gap (**see the answer to** "How does GenBank define a sequence 'gap'?) is located. **If the gap length is estimated**, insert the equivalent number of nnns to represent the gap. If the gap length is unknown, insert 100 n's.

Note: GenBank cannot accept sequences where 50% or more of the submitted sequence is represented by internal Ns (See the answer to "Can I submit a sequence to GenBank that has gaps in it?" for more information about formatting internal Ns).

Below is a IUPAC (International Union of Pure and Applied Chemistry) code table for your reference.

IUPAC-IUB single-letter base codes:

Code	Base Description
G	Guanine

Table continued from previous page.

Code	Base Description
A	Adenine
T	Thymine (Uracil in RNA)
C	Cytosine
R	Purine (A or G)
Y	Pyrimidine (C or T or U)
M	Amino (A or C)
K	Ketone (G or T)
S	Strong interaction (C or G)
W	Weak interaction (A or T)
H	Not-G (A or C or T) H follows G in the alphabet
B	Not-A (C or G or T) B follows A in the alphabet
V	Not-T (not-U) (A or C or G) V follows U in the alphabet
D	Not-C (A or G or T) D follows C in the alphabet
N	Any (A or C or G or T)

Pure and Applied Chemistry **40** (3) , 277 - 331 (1974)

Nomenclature Committee of the International Union of Biochemistry. Ref: Cornish-Bowden, A. Nucl Acid Res 13, 3021-3030 (1985)

Can I submit a sequence to GenBank that has gaps in it? If so, how do I represent the gaps?

For sequences that are from the same organism and individual, and are part of the same gene or locus, but have some sequence missing (like exons of a gene, where the introns are missing), **you'll need to insert a series of nnnns where each** gap (**see the answer to** "How does GenBank define a sequence 'gap'?) is located.

- **If the gap length is estimated**, insert the equivalent number of nnns to represent the gap
- **If the gap length is unknown**, insert a string of 100 nnns to represent the gap
- **Annotate each gap as a misc_feature** (miscellaneous feature) and include a note describing each gap. For gaps of unknown length, be sure to include in your note an explanation that describes the region(s) or feature(s) that is missing (i.e. the missing sequence represented by the nnnns in your gapped submission sequence). **For example**:
- /note="gap, unknown length", intron 2"
- /note="gap, estimated length, ## base pairs"

Use the gap specifications provided in the Sequin Help documentation when you set-up your FASTA-formatted file for importation into Sequin.

BankIt users follow the bulleted points listed above.

Note: GenBank cannot accept sequences where more than 50% of the submitted sequence is gapped (represented by internal nnns).

Sequence Size

Does GenBank have a minimum size requirement for submitted sequences?

GenBank will process nucleotide sequences submissions that are ≥ 200bp in length (Sequences < 200bp are accepted if they represent complete small RNAs or exons).

IUPAC Use

What are the IUPAC codes for nucleotides?

Code	Base Description
G	Guanine
A	Adenine
T	Thymine (Uracil in RNA)
C	Cytosine
R	Purine (A or G)
Y	Pyrimidine (C or T or U)
M	Amino (A or C)
K	Ketone (G or T)
S	Strong interaction (C or G)
W	Weak interaction (A or T)
H	Not-G (A or C or T) H follows G in the alphabet
B	Not-A (C or G or T) B follows A in the alphabet
V	Not-T (not-U) (A or C or G) V follows U in the alphabet
D	Not-C (A or G or T) D follows C in the alphabet
N	Any (A or C or G or T)

Pure and Applied Chemistry **40** (3) , 277 - 331 (1974)

Nomenclature Committee of the International Union of Biochemistry. Ref: Cornish-Bowden, A. Nucl Acid Res 13, 3021-3030 (1985).

Any IUPAC (International Union of Pure and Applied Chemistry) approved single-letter base code for nucleotides, including N, is acceptable for nucleotide sequence data submitted to GenBank.

What are the IUPAC codes for amino acids?

Code	Amino Acid	Code	Amino Acid
A	alanine	P	proline
B	aspartate or asparagine	Q	glutamine
C	cystine	R	arginine
D	aspartate	S	serine
E	glutamate	T	threonine
F	phenylalanine	U	selenocysteine
G	glycine	V	valine
H	histidine	W	tryptophan
I	isoleucine	Y	tyrosine
K	lysine	Z	glutamate or glutamine
L	leucine	X	any amino acid

Table continued from previous page.

Code	Amino Acid	Code	Amino Acid
M	methionine		
N	asparagine		

IUPAC-IUB Joint Commission on Biochemical Nomenclature, Nomenclature and Symbolism for Amino Acids and Peptides section 3AA-1: Names of common α-Amino Acids.

Any IUPAC (International Union of Pure and Applied Chemistry) approved single-letter base code for nucleotides, including X, is acceptable for nucleotide sequence data submitted to GenBank.

Making Tab-delimited Tables

What is a tab-delimited table?

A **tab-delimited table is a table where a single tab keystroke "delimits" (marks the boundary) between one column and the next in a table**.

The format requirements of each tab-delimited table are different, and therefore you should consult the specific table instructions for the resource you are using before you begin your table. Regardless of the type of tab-delimited table you are making, **always follow these rules when making a tab-delimited table**:

- Do **not** use more than one tab keystroke between columns in the table to make the data in the columns align.
- Do **not** use the space bar (in addition to your single tab keystroke) between columns in the table to make the data in the columns align.
- When you save the table:
- **Save the table as plain text (.txt)**.
- Do not save the table as rich text (.rtf) or as a document (.doc)

See the **answer to** "Can you give me step-by-step instructions for making a tab-delimited feature table…", located in this section to see step-by-step **instructions for making a tab-delimited feature table**.

Feature Table

Can you give me step-by-step instructions for making a tab-delimited feature table for my GenBank submission?

A tab-delimited feature table uses a single "Tab" keystroke to delimit (mark the boundary) between one column and the next in a table that contains your feature information.

Follow these instructions to make a tab-delimited feature table:

1. Open a text editor or spreadsheet program:
   a. Microsoft (MS) operating system:
   Open Notepad (Wordpad) or other MS compatible text editor
      - Save document as plain text (.txt).
      - **Do not save** as rich text (.rtf) or as a document (.doc)
   b. Mac operating system:
   Open Textedit, TextWrangler or other Mac compatible text editor.
      - Save document as **plain text (.txt)**.
      - **Do not save** as rich text (.rtf) or as a document (.doc)
   c. Spreadsheet program:
   Open program, enter your data.

- Save document as **plain text (.txt)**.

2. Create a Sequence ID (SeqID) Row.
   The SeqID Row tells our submission system that a new set of features for the SeqID specified in this row will follow. The SeqID Row contains the following:
   - >Features, a space (hit the spacebar on your keyboard once), and the SeqID of the sequence you are annotating. In the example below, eIF4E is the SeqID used in the FASTA file for the sequence:
     `>Features lcl|eIF4E`
3. Hit the Enter key of your keyboard once to go to the next row.
4. Create a Feature Row:
   A Feature Row begins the column portion of the table. The table is composed of five columns (Start, Stop, Feature, Modifier and Modifier value), where each column is separated from the columns beside it by a **single tab keystroke (represented here by <tab>)**.
   a. The Feature Row provides the span (start and stop values) and the type of feature you are supplying for the SeqID indicated in the SeqID Row:
      - SeqID Row: >Features SequenceID
        Feature Row: **Start** <tab> **Stop** <tab> **Feature**
      - Here is an example of how the Feature Row would look in a text editor following the SeqID Row (See Box 1):
   b. Additional intervals (if any) for a particular feature will appear in the rows following the Feature Row, where each interval contained in the feature is represented by its start value and the stop value (span) in its own row.

      Feature Row: **Start value** <tab> **Stop value** <tab> **Feature**
      Interval Row: **Start value** <tab> **Stop value**
      Interval Row: **Start value** <tab> **Stop value**
      Interval Row: **Start value** <tab> **Stop value**

      Here is an example of how Feature Row and its additional intervals would look in a text editor (See Box 2).
5. Hit the Enter key of your keyboard once to go to the next row.
6. Create a Modifier Row:
   A Modifier Row comes at the conclusion of a Feature Row (and any associated Interval Rows), and contains the modifier information for the Feature described in the row(s) above it. The Modifier Row provides the type of modifier as well as the value for that modifier. The Modifier Row begins with three tab keystrokes followed by the modifier name and then by the modifier value.

   If you do not have modifiers to describe the feature provided in the Feature Row, skip down to step 8
   a. This is how you would enter a Modifier Row directly following a Feature Row for a particular SeqID:

      SeqIDRow: >Features SequenceID
      Feature Row: **Start value** <tab> **Stop value** <tab> **Feature**
      Modifier Row: <tab> <tab> <tab> **Modifier** <tab> **Modifier value**

      Here is an example of how the Modifier Row would look row in a text editor when it follows directly after the feature (See Box 3).

b. This is how you would enter a Modifier Row following a Feature Row and its additional intervals:

Feature Row: **Start value** <tab> **Stop value** <tab> **Feature**
Interval Row: **Start value** <tab> **Stop value**
Interval Row: **Start value** <tab> **Stop value**
Interval Row: **Start value** <tab> **Stop value**
Interval Row: **Start value** <tab> **Stop value**
Modifier Row: <tab> <tab> <tab> **Modifier** <tab> **Modifier value**

Here is an example of how the Modifier Row would look in a text editor when it follows a Feature Row and its Interval Rows (See Box 4).

7. Hit the Enter key of your keyboard once to go to the next row.
8. At this point you can do any one of the following:
    - Create another Modifier Row to provide more information for the feature you described in the Feature Row.
    - Create another Feature Row and proceed to describe the intervals and modifiers for this feature using Interval Rows (if any) and Modifier Rows (if any).
    - Create a new SeqID Row, and proceed to describe the features and modifiers for this SeqID using Feature Rows, Interval Rows (if any) and Modifier Rows (if any).
    - Here is how the examples used in the steps above would look together:

SeqIDRow: >**Features SequenceID**
Feature Row: **Start value** <tab> **Stop value** <tab> **Feature**
Modifier Row: <tab> <tab> <tab> **Modifier** <tab> **Modifier value**
Feature Row: **Start value** <tab> **Stop value** <tab> **Feature**
Interval Row: **Start value** <tab> **Stop value**
Interval Row: **Start value** <tab> **Stop value**
Interval Row: **Start value** <tab> **Stop value**
Interval Row: **Start value** <tab> **Stop value**
Modifier Row: <tab> <tab> <tab> **Modifier** <tab> **Modifier value**

Here is how the examples used in the steps above would look together in a text editor (See Box 5):

9. Once your table is imported into Sequin (or BankIt), Sequin/BankIt will recognize the SeqIDs in your table, and will automatically assign and place the appropriate features and their modifiers on each sequence in your set

When you create your feature table, **always remember the following**:

- When you make your tables:
    - Do **not** use more than one tab keystroke between columns in the table to make the data in the columns align.
    - Do **not** use the space bar (in addition to your single tab keystroke) between columns in the table to make the data in the columns align.
- When you save the table:
    - Save the table as **plain text (.txt)**.
    - Do not save the table as rich text (.rtf) or as a document (.doc)

You can see the complete feature table for lcl|eIF4E (the example used above) in the Sequin Quick Guide, and you can also find an additional example of a more complex feature table in the "Submission of Annotation using a Table" page (scroll down to see the example in Fig. 1).

Box 1.

```
>Features lcl|eIF4E
80      2881    gene
```

Box 2.

```
201     224     CDS
1550    1920
1986    2085
2317    2404
```

Box 3.

```
>Features lcl|eIF4E
80      2881    gene
                        gene    eIF4E
```

Box 4.

```
201     224     CDS
1550    1920
1986    2085
2317    2404
2466    2629
                            product     eukaryotic initiation factor 4E-II
```

Box 5.

```
>Features lcl|eIF4E
80      2881    gene
                        gene    eIF4E
201     224     CDS
1550    1920
1986    2085
2317    2404
2466    2629
                        product  eukaryotic initiation factor 4E-II
```

Source Modifier Table

Can you give me step-by-step instructions for making a tab-delimited source modifier table for my GenBank submission?

Note: Due to a technical issue with the right margin that cannot be fixed, the example lines in this "step-by-step" for the source modifier table have been broken into two lines, but it is important that when you enter lines like these, they should be in a single line without breaks.

A tab-delimited source modifier table uses a single "Tab" keystroke to delimit (mark the boundary) between one column and the next in a table that contains your source modifier information.

Follow these instructions to make a tab-delimited source modifier table:

1. Open a text editor or spreadsheet program:
   a. Microsoft (MS) operating system:
   Open Notepad (Wordpad) or other MS compatible text editor.
      - Save document as plain **text (.txt)**.
      - Do not save as rich text (.rtf) or as a document (.doc).
   b. Mac operating system:
   Open Textedit, TextWrangler or other Mac compatible text editor.
      - Save document as plain **text (.txt)**.
      - Do not save as rich text (.rtf) or as a document (.doc).
   c. Spreadsheet program:
   Open program, enter your data.
      - Save document as **plain text (.txt)**.
2. **Create the Column Label Row**:
   - The First column always lists the Sequence IDs; each subsequent column in the table lists a different source modifier that will be applied. You will find a comprehensive list of source modifiers online.
      a. Type this label: **Sequence_ID** as the first entry in the Column Label Row. The Sequence_ID must be the same as that used to identify each sequence in your nucleotide FASTA file.
      b. Separate the Sequence_ID label by a single tab key stroke from the next column, which will be a source modifier label. Add as many source modifier labels to this row as you need, each separated from the next by a single tab keystroke. The column labels can be in whatever order you want so long as the Sequence_ID label starts the Column Label Row.
         i. This is how you would enter a Column Label Row:

            Column Label Row:
            Sequence_ID<tab>Specimen_voucher<tab>Collected_by<tab>
            Collection_date<tab>Country<tab>Identified_by <tab> Lat_Lon
         ii. Here is an example of how the Column Label Row would look in a text editor (see Box 6)
3. **Hit the Enter key of your keyboard** once to go to the next row.
4. **Create a Source Modifier Row** for your first sequence:
   a. Enter the sequence ID of your first sequence in the first column of the table. Enter a single tab keystroke, followed by the source modifier value for the source modifier column that follows the Sequence_ID column.
   b. Enter another tab keystroke followed by the source modifier value for the second source modifier column that follows the Sequence_ID column. Continue to enter the value data for each source modifier label (each value separated by a single tab keystroke from the next) until all the source modifier data for the first SeqID has been entered.
      i. This is how you would enter a Source Modifier Row following a Column Label Row:

         Column Label Row: Sequence_ID<tab>Specimen_voucher<tab>Collected_by<tab>
         Collection_date<tab>Country<tab>Identified_by <tab>Lat_Lon
         Source Modifier Row 1: Seq1<tab>MKP 334<tab>C. Grant<tab>31-Jan-2001 USA
         <tab> C. Grant<tab>13.57N 24.68 W
      ii. Here is an example of how the Column Label Row and Source Modifier Row would look in a text editor (see Box 7).

5. **Create a Source Modifier Row for your second sequence:**
   a. Enter the sequence ID of your second sequence in the first column of the table. Enter a single tab keystroke, followed by the source modifier value for the source modifier column that follows the Sequence_ID column.
   b. Enter another tab keystroke followed by the source modifier value for the second column that follows the Sequence_ID column. Continue to enter data for each source modifier label (each separated by a single tab keystroke from the next) until all the source modifier data for the second SeqID has been entered.
      i. This is how you would enter a second Source Modifier Row following a Column Label Row and a preceding Source Modifier Row:

         Column Label Row: Sequence_ID<tab>Specimen_voucher<tab>Collected_by<tab> Collection_date<tab>Country<tab>Identified_by <tab>Lat_Lon
         Source Modifier Row 1: Seq1<tab> MKP 334<tab>C. Grant<tab>31-Jan-2001 USA <tab> C. Grant<tab> 13.57N 24.68 W
         Source Modifier Row2: Seq2<tab>MKP 1230<tab>S. Tracy<tab> 28-Feb-2002<tab>Slovakia <tab>C. Grant<tab>13.24 N 24.35 W
      ii. Here is an example of how the Column Label Row and Source Modifier Rows would look in a text editor (See Box 8).
         - Because you are using Tabs to separate columns, each Source Modifer Row column may not line up with the Column Label Row columns or with other Source Modifier Row columns. This is normal and valid, as long as you are using only Tabs between each column.
         - Do **not** use more than one tab keystroke between columns in the table to make the data in the columns align.
         - Do **not** use the space bar (in addition to your single tab keystroke) between columns in the table to make the data in the columns align.
6. Continue to add Source Modifier Rows for each of your remaining sequences.
   a. This is how you would enter an additional Source Modifier Rows following a Column Label Row and preceding Source Modifier Rows:

      Column Label Row: Sequence_ID<tab>Specimen_voucher <tab>Collected_by<tab> Collection_date<tab>Country<tab> Identified_by<tab>Lat_Lon
      Source Modifier Row 1: Seq1<tab>MKP 334 <tab>C. Grant<tab>31-Jan-2001<tab>USA<tab> C. Grant<tab>13.57 N 24.68 W
      Source Modifier Row 2: Seq2<tab>MKP 1230 <tab>S. Tracy<tab>28-Feb-2002<tab>Slovakia <tab>C. Grant<tab>13.24 N 24.35 W
      Source Modifier Row 3: Seq3<tab>1B-2526<tab>A. Gardner<tab> 16-Apr-2001France<tab>C. Grant <tab>43.21 N 56.78 W
      Source Modifier Row 4: Seq4<tab>WBM 86-64<tab>F. McMurray<tab> 26-May-2002<tab>Germany<tab>C. Grant<tab>45.32 N 21.34 E
      Source Modifier Row 5: Seq5<tab> 1B-2518<tab>V. Leigh<tab>13-Jun- 2003<tab>Brazil <tab>V. Leigh<tab>46.80 N 13.57 E
   b. Here is an example of how the Column Label Row and additional Source Modifier Rows would look in a text editor (See Box 9.)
7. Once your table is imported into Sequin (or BankIt), Sequin/BankIt will recognize the SeqIDs in your table, and will automatically assign and place the appropriate source modifiers on each sequence in your set.
8. When you create your Source Modifier table, **always remember the following**:

- Do **not** use more than one tab keystroke between columns in the table to make the data in the columns align.
- Do **not** use the space bar (in addition to your single tab keystroke) between columns in the table to make the data in the columns align.
- Each Sequence ID (SeqID) can appear **only once** in a source modifier table.
- When you save the table:
- Save the table as **plain text** (.txt).
- Do **not** save the table as rich text (.rtf) or as a document (.doc).

Please see the BankIt Submission Help documentation for further information on creating source modifier tables.

Box 6.

```
Sequence_ID    Specimen_voucher    Collected_by    Collection_date    Country
Identified_by     Lat_Lon
```

Box 7.

```
Sequence_ID   Specimen_voucher   Collected_by   Collection_date    Country
Identified_by   Lat_Lon
Seq1        MKP 334       C. Grant        31-Jan-2001       USA      C. Grant       13.57
N 24.68 W
```

Box 8.

```
Sequence_ID   Specimen_voucher   Collected_by   Collection_date   Country
Identified_by   Lat_Lon
Seq1     MKP 334     C. Grant     31-Jan-2001     USA     C. Grant     13.57 N 24.68 W
Seq2     MKP 1230     S. Tracy     28-Feb-2002     Slovakia     C. Grant     13.24 N
24.35 W
```

Box 9.

```
        Sequence_ID   Specimen_voucher   Collected_by   Collection_date   Country
Identified_by   Lat_Lon
        Seq1     MKP 334     C. Grant     31-Jan-2001     USA     C. Grant     13.57 N
24.68 W
        Seq2     MKP 1230     S. Tracy     28-Feb-2002     Slovakia     C. Grant
13.24 N 24.35 W
        Seq3     1B-2526     A. Gardner     16-Apr-2001     France     C. Grant     43.21
N 56.78 W
        Seq4     WBM 86-64     F. McMurray     26-May-2002     Germany     C. Grant
45.32 N 21.34 E
        Seq5     1B-2518     V. Leigh     13-Jun-2003     Brazil     V. Leigh     46.80 N
13.57 E
```

Annotating your Sequence for Submission

Created: April 6, 2011; Updated: November 3, 2014.

Why Should I Add Features to my Sequence?

What do you mean by feature annotation and why do I need to annotate my sequences?

Feature annotation is the addition of biological features such as genes and associated coding regions, structural RNA, variation information, exon, introns, etc. to your submitted sequence. The annotation should include the location of the feature (start and stop) and a description of the feature.

The addition of feature annotation to your sequence submission:

- Improves the quality of your submission
- Increases the efficiency with which your submitted sequences are processed by members of the GenBank staff
- Is of far greater use to the scientific community than sequence data alone.

Adding feature annotation will frequently provide an additional tool for reviewing the quality of primary nucleotide sequence data:

For example, annotating protein-coding regions will highlight potential errors in the nucleotide sequence, such as insertion/deletions (in/dels) or improper or uncertain base calls that result from the sequencing reads.

See an alphabetic list **of available Features** in the Sequin Help documentation (These Features can be used in both Sequin and BankIt).

See the "Annotation using BankIt" and "Annotation using Sequin" sections of this Quick Start for information about how to annotate your sequences.

Can I submit a sequence without annotating it?

You must provide some type of annotation with your sequence such as:

- Coding Sequence (CDS), including nucleotide spans and reading frame. Using this information, our software will add the amino acid translations for you.
- structural RNAs such as rRNAs, tRNAs, misc_RNAs (miscellaneous RNAs), with nucleotide spans (if known)
- features which may describe your sequence, such as repeat regions, UTRs, promoters with nucleotide spans

The addition of feature annotation to your sequence submission:

- Improves the quality of your submission
- Increases the efficiency with which your submitted sequences are processed by members of the GenBank staff
- Is of far greater use to the scientific community than sequence data alone.

Adding feature annotation will also provide an additional tool for reviewing the quality of primary nucleotide sequence data.

For example, annotating protein-coding regions will frequently highlight potential errors in the nucleotide sequence, such as insertion/deletions (in/dels) or improper or uncertain base calls that result from the sequencing reads.

See an alphabetic list **of available Features** in the Sequin Help documentation (These Features can be used in both Sequin and BankIt).

See the "Annotation using BankIt" and "Annotation using Sequin" sections of this Quick Start for information about how to annotate your sequences.

Feature Annotation Using BankIt

How do I annotate features in my submission using BankIt?

1 **At step 8 ("Features") of the BankIt submission process, you will choose between:**
   a. Uploading a 5-column, tab-delimited table file containing your sequence features (select: "File" button)

   OR
   b. Picking feature categories and feature types for your sequence from a list provided in an online BankIt form (select: "Form" button).

Benefits of using File data upload:

- Good for different features on multiple sequences
- Helpful for adding many multiple features on a single sequence or on multiple sequences
- Uses the five-column, tab-delimited feature table format
- Multiple tables can be uploaded in a single file

Benefits of using Form input:

- Good for a single feature or a few features applied to a single sequence
- Good for applying a single feature to all sequences in a set or batch, or for applying a few of the same features to all sequences in a set or batch
- Features can be added across an entire sequence or by specific intervals within a sequence
- One or more modifiers can be chosen to apply to each feature

2. **If you select the File upload button**:
   a. **Click the "Browse" button and select the feature table .txt** file you would like to upload.
   b. **Click the "Upload File" button** and upload your file.
   c. **Find the "Current Features" section of the Features (Overview) page**, where you will see a list of the features created from the table you just uploaded. Next to each feature on the list are buttons that allow you to either edit a feature or remove it entirely.
   d. Scroll to the bottom of the "Current Features" section and **click "Continue" to go to the next step of the submission process** once the features have been entered to your satisfaction.
   e. Features can be edited or deleted before continuing to the Review and Submit steps.
3. **If you select the "Form" button:**
   a. **Select one category from the five general feature categories presented:**
      - CDS/gene/mRNA
      - structural RNAs
      - Gene
      - Repeat Region (for simple repeats, mobile elements, and satellites)
      - Other (e.g.: D-loop, misc_feature, polyA_site, variation, etc.)
   b. **Select the appropriate feature type if presented with a choice** after selecting the feature category.

c. **Click the "Add" button**. On the new page that appears, provide the specific details for the feature (e.g. nucleotide intervals, protein name, rRNA name, gene name, etc.).
d. **Click the "Accept" button** at the bottom of the page.
e. **On the "Features (Overview)" page, find the "Current Features" section**. You will see a list of the features created using the data you provided. Next to each feature on the list are buttons that allow you to either edit a feature or remove it entirely before continuing to the Review and Submit steps.
f. **Go to the bottom of the "Current Features" section of the page, and click "Continue"** to go to the next step of the submission process once the features have been entered to your satisfaction.

Annotation of Coding Regions using BankIt

How do I add annotation for coding regions in my submission using BankIt?

The easiest way to add annotation for coding regions in your submission is to:

- Provide the coding region spans when prompted by your submission program to do so; **the submission toolwill automatically translate the feature span for you.** (we prefer this means of generating the translation)

but, you can also:

- Import the amino acid sequence and ask the submission program to predict the coding region spans for you.
  - Once you import the protein sequence, BankIt will process the translation for you, and will inform you if there are errors in the translated intervals. If you are notified of errors, you will then be prompted to make corrections to your sequence before proceeding with your submission.
  - Although we will accept this means of generating a translation, **we would prefer that you submit the span information**.

Do I have to submit the translated sequence when I annotate my submission using BankIt?

Generally, **there is no need to provide the translations yourself since**:

- **If you provide the span information** for the feature when prompted to do so during your submission**,** the submission tool will automatically translate the feature span for you.
  - This is easiest way to add annotation for coding regions in your submission.
  - **We prefer this means of generating the translation.**
- **If you do not have span information**, you can import the protein and have the submission program translate the protein to get the information for the feature you wish to annotate.
  - Once you import the protein sequence, BankIt will process the translation for you, and will inform you if there are errors in the translated intervals. If you are notified of errors, you will then be prompted to make corrections to your sequence before proceeding with your submission.
  - Although we will accept this means of generating a translation, **we would prefer that you submit the span information**.

Feature Annotation Using Sequin

How do I annotate features in my submission using Sequin?

As Sequin has a number of annotation options, the answer to this question depends on the nature of the annotation you wish to add to your sequence:

- **If you are submitting a single sequence and wish to annotate a singlefeatureto it:**
  You can either enter them when prompted by Sequin during the submission process, or you can use the "Record Viewer Annotate Menu", which will provide you with a list of annotation options.

For more information on how to use the Record Viewer Annotate Menu, see the Features section of the Sequin Help Documentation.

- **If you are submitting a set of similar sequences, and want to annotate the same feature across the entire span of each:**
  Use the Batch Feature Apply option *ifthe feature you wish to annotate spans the entire nucleotide sequence of each member of the set. You cannot annotate specific nucleotide locations using this option.*

For more information about the Batch Feature Apply option, see the Annotate Menu section of the Sequin Help Documentation.

- **If you are submitting an aligned set of sequences, and want to annotate the same feature that you added to one sequence to all the other sequences within the set:**
  Use the Feature Propagate option. For more information about the Feature Propagate option, see the Feature Propagate section of the Sequin Help Documentation.
- **If you are submitting an aligned set of sequences, and want to annotatefeatures to each sequence in the set:**
  Use the Alignment Assistant. For more information about the Alignment Assistant, see the Alignment Assistant section of the Sequin Help Documentation.
- **If you wish to annotate features to sequence records in Sequin, but prefer not to use the options mentioned above:**
  You can create a five-column, tab-delimited feature table and import it into Sequin.

See the step-by-step instructions for making a tab-delimited table in this Quick Start.

See the "Submission of Annotation Using a Table" page of the Sequin help documentation for additional information about the use of annotation tables in Sequin.

Annotation of Coding Regions using Sequin

How do I add annotation for coding regions in my submission using Sequin?

The easiest way to add annotation for coding regions in your submission is to:

- **Provide the coding region spans** when prompted by your submission program to do so; the submission tool will automatically translate the feature span for you. (we prefer this means of generating the translation)

but, you can also:

- Import the amino acid sequence and ask the submission program to predict the coding region spans for you.
  - If you choose to proceed using this means of generating a translation, **you must validate your submission to make sure the predicted spans are correct before you submit**. Please check any validation warnings generated by Sequin, and correct any errors that you find.
  - Although we will accept this means of generating a translation, we would prefer that you submit the span information.

Do I have to submit the translated sequence when I annotate my submission using Sequin?

Generally, **there is no need to provide the translations yourself since**:

- **If you provide the span information** for the feature when prompted to do so during your submission**,** the submission tool will automatically translate the feature span for you
  - This is easiest way to add annotation for coding regions in your submission.
  - **We prefer this means of generating the translation.**
- **If you do not have span information**, you can import the protein and have the submission program translate the protein to get the information for the feature you wish to annotate.
  - If you choose to proceed using this means of generating a translation, **you must validate your submission to make sure the predicted spans are correct before you submit**. Please check any validation warnings generated by Sequin, and correct any errors that you find.
  - Although we will accept this means of generating a translation, we would prefer that you submit the span information.

Providing Source Information in your Submission

Created: April 6, 2011; Updated: November 3, 2014.

Source information for Samples Collected in the Field

What is an Isolation_Source?

I have isolated sequences from samples I obtained in the field, and have been asked to provide an isolation_source — what is an isolation_source?

The isolation_source is a modifier that describes the physical, environmental and/or local geographical source of the biological sample from which a sequence was derived (e.g. soil, sediment, ocean water, lake water, forest debris, soil from outside a specific chemical factory, gasoline polluted soil, etc.)

Country and Latitude/Longitude Information

How specific should my latitude longitude (lat_lon) information be?

Provide the lat_lon in decimal degrees that include compass direction (e.g. 39.7 N 42.1 W). Also provide the country of origin for your sample if you have it.

Since I don't know the latitude and longitude (lat_lon) of my samples, should I give the latitude/longitude based on the lat_lon of a near-by city or landmark?

- **Do not provide lat_lon information if** you have to determine the latitude and longitude of the sample site (based on a nearby city or landmark) after the sample was collected.

  If you do not have the lat_lon information for your sample, provide the country of origin for your sample. Include the specific locality where the sample was obtained (if known), using the following format:

  /country=”country:sub_region”
  For example:
  /country="Canada:Vancouver"
  Or
  /country="USA: Bethesda, MD"
- **Provide lat_lon information only if** you recorded it at the time you collected the sample. Also provide the country of origin for your sample if you have it.

Additional Collection Information

How do I add (extensive/detailed/complicated) isolation, location, or other information to my submissions for the organisms from which I isolated my sequences?

Detailed or extensive information describing the location where your organisms were isolated or detailed organism descriptions can be included in the following source modifiers:

- /authority

  The /authority modifier should include the complete name of the organism, followed by the authority information.

  For example:

/authority=Elekmania picardae (Krug & Urb.) B.
/authority=Avena sativa L.

- /collected_by

  The /collected_by modifier should report the name(s) of the specific person(s) who collected the organism from which the sequence was obtained.

  For example:
  /collected_by=Fred MacMurray
  /collected_by=A. Hitchcock and F. Zeferelli
- /collection_date

  The /collection_date modifier must be in the format DD-Mon-YYYY or Mon-YYYY or YYYY.

  For example:
  /collection_date=30-Sep-2008
  /collection_date=Sep-2008
  /collection_date=2008
- /country

  The /country modifier can also include province, state, region, oceans, or other locality names. The name of the country (or ocean) must be provided first, followed by a colon (:) before the additional location information.

  For example:
  /country=USA: Lancaster County, PA
  /country=Canada: SW coast of Newfoundland
  /country=USA: Syracuse State Park in upstate New York
  /country=Atlantic Ocean: 24.5 miles east of Bermuda
  /country=Pacific Ocean: Stubing Marine Station

  You can find a list of INSDC approved country names online at NCBI.
- /identifed_by

  The /identifed_by modifier reports the name(s) of the specific person(s) who identified the TAXONOMY of the organism from which the sequence was obtained. This does not mean the person(s) in the laboratory who identified the submitted sample.

  For example:
  /identified_by=James Cagney
  /identified_by=K. Hepburn and S. Poitier
- /isolation_source

  The /isolation_source modifier describes the physical, environmental, and/or local geographic location where the organism was isolated.

  For example:
  /isolation_source=cow rumen
  /isolation_source=abandoned silver mine 20 miles NW of Las Vegas

/isolation_source=roadkill on Old Sulphur Mill Road
/isolation_source=activated sludge from bioreactor
/isolation_source=runoff from chicken farm

- /lat-lon

  The /lat-lon modifier should be in a decimal degree format using single letters that denote direction and should report the latitude and longitude measured at the site and time of collection.

  For example:
  /lat_lon=47.68 N 33.75 W
  /lat_lon=28.82 S 12.50 E

Note:

- The latitude and longitude **cannotbe estimated** from a map or GPS device **after** the collection is complete.
- The specific country (or ocean) should also be reported as a/country modifier.

Additional organism metadata that do not fit into any of the available modifiers may be appropriate in a Structured Comment.

Eukaryotic Source Material

What kind of source information do I provide with a Eukaryotic Set submission?

- **If the source material originated in a museum or other reference collection, provide the following information only if** the sequence you are submitting was obtained from a sample that you retrieved directly from the indicated museum/collection, or the sequence was obtained from a sample that you deposited in the indicated museum/collection:
  - The specimen voucher number for each different source material used**.** Specimen vouchers provide a means to verify the identity of a taxon and are a source for additional molecular analyses.
  - **See the "**Museum/Reference Collection Source Material**" section** to see the types of information accepted for museum/reference collection source materials).
- **If your source material does not come from a museum or other reference collection, and has no specimen voucher number**, provide any of the following information about your source material(s):
  - Cultivar, strain, isolate, or breed
  - Germplasm, seed, or stock center accession number (use biomaterial modifier)
  - Collection number, locality, and/or date

Rat/Mouse Source Material

What kind of source information do I provide with a set of mouse/rat sequences?

In addition to the information for eukaryotic source material(s), provide the mouse/rat strain name for each mouse or rat strain submitted. If you do not know the strain name, please tell us at the time of your submission.

Bacterial and Archaeal Source Material

What are strain identifiers? Why do I need to provide one for each cultured bacterial/archaeal sequence submission?

Strain identifiers serve to distinguish your culture from other related isolates of the same species obtained in your lab or elsewhere.

Note: your isolates do not need to be deposited in a culture collection in order for you to create strain identifiers for them.

Each cultured bacterial/archaeal sequence you submit should have a unique strain identifier associated with it.

- If your culture comes from a culture collection and has an established identifier, use it as the strain identifier.
- If your culture does not come from a culture collection, you can create **a strain identifier by using anything that identifies the particular culture from which the sequence was obtained:**
  - Any distinguishing identifier you use in the laboratory
  - A string of numbers and/or letters

Since the species name of your bacterial/archaeal isolate alone does not uniquely identify a particular culture, you must provide an identifier for a particular culture of the bacterial/archaeal isolate.

If a species has not yet been assigned to your isolate, you still must provide an identifier for it using the suggestions for creating a strain identifier mentioned above.

How do I provide unique source information for my bacteria/archaea submission?

If you are submitting a number of different sequences isolated from different strains/isolates/clones, provide the information as a spreadsheet or tab-delimited table:

SeqID	strain
Seq01	ABC
Seq02	CBS 235

etc.

If the same sequence was found in separate strains/isolates/clones, create an additional sequence submission for each source type and submit the group of sequences together.

Or you can provide a tab-delimited table giving a single source in the modifier column (e.g. strain) for the sequence, and then list in a note within the table any additional sources where you found the sequence:

SeqID	strain	note
Seq01	ABC	identical sequence found in strains DEF and GHI
Seq02	CBS 235	identical sequence found in strains CBS 236 and CBS 237

etc.

If I'm submitting sequence obtained from uncultured bacteria/archaea, what descriptive information do I need to provide about the source?

If you are submitting sequence from an uncultured source, in addition to the information presented in the question/answer unit about bacterial/archaeal genomic sequence submissions, identify the submission source material as uncultured, and provide the following:

Details describing the isolation source (environmental conditions) where the bacteria/archaea was isolated and a unique clone identifier. If you are submitting sequences isolated from multiple conditions and/or locations, provide the information as a spreadsheet or tab-delimited table:

SeqID	isolation source	clone
Seq01	soil	Qx27
Seq02	ocean water	Qy28

etc.

If the same sequence was found in separate isolation sources, create an additional sequence submission for each source type and submit the group of sequences together.

Or you can provide a tab-delimited table giving a single source in the modifier column (e.g. environment) for the sequence, and then list in a note within the table any additional sources where you found the sequence:

SeqID	isolation source	note
Seq11	soil	identical sequence found in pond water and tree bark
Seq12	ocean water	identical sequence found in sand and in samples collected from surface of coastal boulders

etc.

Viral Source Material

What kind of source information do I include in a viral sequence submission?

Provide the following information with your viral sequence submission in a tab-delimited source table:

- Strain or Isolate
- Serotype or Genotype, if appropriate
- Host
- Country
- Collection_date

See an online example of the annotation of a viral sequence submission.

Museum/Reference Collection Source Material

What kind of source information do I include in submissions whose sequence is extracted from specimens obtained from museums or from other reference collections?

There are three different source modifiers that define the type of information accepted for museum/reference collection source materials:

- culture_collection
- specimen_voucher
- bio_material

The type of information to submit with your sequence depends upon which of these three modifiers best describes the source material from which you extracted your sequence.

Each of these three modifiers is defined below. Select the modifier below that best describes the source material you have. The specific information you need to provide and examples of that information follow the definition of each source type.

1. **/culture_collection**:
    - This modifier is used to annotate the following source material types:
    - Live microbial and viral cultures

- Cell lines that have been deposited in curated culture collections
- **Provide the following information only if** the sequence you are submitting was obtained from a sample you retrieved directly from the indicated culture collection, or the sequence was obtained from a sample that you deposited in the indicated culture collection:
- The institution code for the institution where the culture is housed (mandatory).
- The identifier of the culture from which the nucleic acid sequenced was obtained (mandatory).
- Optional collection code.
  A searchable database of institution and collection codes is currently being developed.
  Note: The institution-code (and optional collection-code) must be taken from the INSDC controlled vocabulary (preselected and predefined authorized terms) that denote the institution/collection where the culture is maintained.
- **The format of the /culture_collection information you provide** can be either of the following:

  ```
  /culture_collection=institution-code:specimen_id
  /culture_collection=institution-code:collection-
  code:specimen_id
  ```
- **Example of the /culture_collection information**:

  ```
  /culture_collection="ATCC:26370"
  ```

  If you annotate a sequence **with more than one culture_collection modifier**, this indicates that the sequence was obtained from a sample that was deposited (by the submitter or a collaborator) in more than one culture collection.

  Microbial cultures in personal or laboratory collections should be annotated using strain modifiers

2. **/specimen_voucher**:
   - This modifier is used to annotate the following source material:
   - **A physical specimen that remains after the sequence has been obtained**.
   - Ideally the specimen(s) is/are housed in a curated museum, herbarium, or frozen tissue collection, but it/they can be housed in a personal or laboratory collection as well. If the specimen was destroyed in the process of sequencing, electronic images (e-vouchers) are an adequate substitute for a specimen voucher that identifies physical remains.
   - **Provide the following information only if** the sequence you are submitting was obtained from a sample you retrieved directly from the indicated museum/collection, or the sequence was obtained from a sample that you deposited in the indicated museum/collection:
   - The unique identifier of the specimen from which the nucleic acid sequenced was obtained. (mandatory)
   - The institution code for the institution where the specimen is housed. Omit the institution code if the specimen comes from a personal collection.
   - Optional collection code.
     A searchable database of institution and collection codes is currently being developed.
     Note: The institution-code (and optional collection-code) must be taken from the INSDC controlled vocabulary (preselected and predefined authorized terms) that denote the institution/collection where the specimen resides.
   - **The format of the /specimen_voucher information you provide** can be any of the following:

     ```
     /specimen_voucher=specimen_id
     ```

```
/specimen_voucher=institution-code:specimen_id
/specimen_voucher=institution-code:collection-code:specimen_id
```

- **Examples of the /specimen_voucher information:**

```
/specimen_voucher="UAM:Mamm:52179"
/specimen_voucher="AMCC:101706"
/specimen_voucher="USNM:field series 8798"
/specimen_voucher="personal:Dan Janzen:99-SRNP-2003"
/specimen_voucher="99-SRNP-2003"
```

3. **/bio_material**:
   - This modifier is used to annotate source material in biological collections that do not fit into either the /specimen_voucher or the /culture_collection modifier categories:
   - Physical specimens from **zoos**
   - Physical specimens from **aquaria**
   - Physical specimens from **stock centers**
   - Physical specimens from **seed banks**
   - Physical specimens from **germplasm repositories**
   - Physical specimens from **DNA banks**
   - **Provide the following information only if** the sequence you are was obtained from a sample you retrieved directly from the indicated collection, or the sequence was obtained from a sample that you deposited in the indicated collection:
   - The identifier of the biological material from which the nucleic acid sequenced was obtained (mandatory)
   - Optional institution code
   - Optional collection code. If you decide to include the collection code, you must also provide the institution code.

A searchable database of institution and collection codes is currently being developed.

Note: The institution-code (and optional collection-code) must be taken from the INSDC controlled vocabulary (preselected and predefined authorized terms) that denote the institution/collection where the specimen resides.

- **The format of the /bio_material information you provide** can be any of the following:

```
/bio_material= =specimen_id
/bio_material= =institution-code:specimen_id
/bio_material= =institution-code:collection-code:specimen_id
```

- **Example:**

```
/bio_material="CGC:CB3912" .
```

How to Describe Unknown Source Material in Your Submission

Can I use the word "unknown" if I don't know the organism from which a sequence came?

GenBank cannot process your sequence and assign an accession number to it if the source material is simply described as "unknown".

Different source materials are described below. Read each description carefully; once you have found the description that best describes the source material you have, the information you will need to provide for the source material follows.

A. **Were the sequences you want to submit derived from**:
   1. **Pure culture?**
      (a culture that contains only one microbial species),

      If so, provide the taxonomic lineage as far as you have determined it, and include the strain identifier. We will accept any of the following for a cultured organism: (See Box 10)
   2. Enrichment culture?
      (use of selective culture media to enrich for a set of microorganisms with a particular phenotypic property, resulting in a partially purified, mixed culture)

      If so, provide the taxonomic lineage as far as you have determined it and include the clone identifier. We will accept either of the following for organisms obtained from an enrichment culture: (See Box 11)

B. **Were the sequences you want to submit extracted from**:
   1. Bulk environmental DNA (using universal primers)?
      DNA that is PCR-amplified directly from source/host DNA (e.g. soil, ocean water, etc.) using universal primers

      If so, provide the taxonomic lineage as far as you have determined it and include the clone identifier in the appropriate source modifier. Do not include the clone within the organism name itself. You must also include the isolation source modifier, within which you will give the specific environmental conditions from which the sample was isolated (soil, ocean water,etc.). We will accept any of the following for the culture lineage: (See Box 12)
   2. Bulk environmental DNA (using species-specific primers, not gene-specific primers)?
      DNA that is PCR-amplified directly from source/host DNA (e.g. soil, ocean water, etc.) using species-specific primers.

      If so, provide the full binomial (genus species) name and include the isolate identifier in the appropriate source modifier. Do not include the isolate within the organism itself. You must also include the isolation source modifier giving the specific environmental conditions from which the sample was isolated (soil, ocean water, etc). Please include a note indicating that the sequence was amplified with species-specific primers. (See Box 13).

Is it OK for me to use BLAST results to identify the source organism from which I isolated my sequences?

Use BLAST scores only as a rough guide to the identity of the source organisms from which sequences were derived.

Assign source organisms to the same taxonomic rank as the best BLAST hits only at the genus level or higher, depending on the consistency among the best hits:

- Assign the source organism to a single genus if the best scores involve taxa from that genus
- Name the source organism using the next highest rank (family, order, class, or phylum) that includes all the best hits if there is inconsistency among the best hits as to the genus identity
- The source organism should be assigned the name 'bacterium <strain>'(for domain Bacteria) OR 'archaeon <strain>' (for domain Archaea) if there is uncertainty as to the proper phylum

Examine the "taxonomy reports" that are provided with the BLAST results. The initial report presents the information in the definition field of the associated sequence records.

While this field should be updated if there is a change in the taxonomic name or lineage of the source organism in that record, the individual, complete sequence record should be examined to be certain that the proper taxonomic information is utilized in the definition field.

Note: species names can be assigned to source organisms only if species-specific primers were used during amplification. Otherwise, you must use genus level or higher ranks to name the organism.

Box 10.

Format	Example
Genus species	Escherichia coli
Genus sp. strain identifier	Escherichia sp. 1234
Family bacterium strain identifier	Enterobacteriaceae bacterium 1234

Box 11.

Format	Example
Genus sp. enrichment culture clone 1234 clone identifier	Escherichia sp. enrichment culture clone
Family bacterium enrichment culture clone culture clone 1234 clone identifier	Enterobacteriaceae bacterium enrichment

Box 12.

Format	Example
uncultured Genus sp.	uncultured Escherichia sp.
uncultured Family bacterium	uncultured Enterobacteriaceae bacterium
uncultured Kingdom	uncultured bacterium

Box 13.

Format	Example
Full binomial (genus species) name	Escherichia coli

Setting Release Dates for your Submission

Created: April 6, 2011; Updated: November 3, 2014.

Can I submit now and have my sequences withheld from the database until a later date?

As you prepare your submission using Sequin or BankIt, **you will be allowed to specify a future release date for the sequence(s) you are submitting**. GenBank will hold your sequence(s) from public view until this date or when the accession numbers or sequence data are published, whichever is first.

Can I ask for an extension of a release date? The paper describing the sequence(s) I submitted will not be published by the release date I specified in my submission.

If the paper describing your sequence will not be published by the release date specified in your submission, **email GenBank at gb-admin@ncbi.nlm.nih.gov and request an extension of the release date** for the accession number(s) involved. Provide a specific release date; we cannot withhold a sequence indefinitely pending publication.

Note: A request for the extension of a release date must be submitted to GenBank well in advance of the original release date specified in your submission.

If the paper describing my sequence is accepted before the release date I provide, how do I ask GenBank to release the sequence(s)?

To request the release of a sequence record prior to the release date specified in your submission, contact the GenBank annotation staff at gb-admin@ncbi.nlm.nih.gov. Include in your request the accession number or range of accessions that you wish to have released and the relevant publication details.

I found a GenBank accession number in a publication, but I can't find the corresponding sequence for it in GenBank. How do I get the sequence released so I can see it?

In order to release the accession(s) you need to public view, **we will need to independently verify that the accessions in question have indeed been published**. What you can do to help us do this is to send either the:

- The citation for the article in which the accession appears. Please include a PDF if possible.

OR

- The PubMed ID (PMID) number for the article in which the accession appears to info@ncbi.nlm.nih.gov

.

Sending your Submission to GenBank

Created: April 6, 2011; Updated: November 3, 2014.

Problems Sending Files by email

I have tried to email my large Sequin file and it will not go through. Do you have a size limit? How do I transmit the file to you?

We don't have a maximum size limit. However, as many email systems will truncate large messages, we suggest you send us your file(s) using SequinMacroSend if you are having trouble sending us a file by email.

SequinMacroSend was designed for uploading large .sqn files directly to GenBank using a web-interface rather than sending large email attachments.

Fill in the information on the SequinMacroSend **form** (the form is located at the bottom of the SequinMacroSend page), **upload your prepared .sqn file, and the submission will be sent directly to the GenBank submission staff**.

Submission Processing

Created: April 6, 2011; Updated: November 3, 2014.

Time Required to Process Submission

How long will it take for me to get my accession number once I've submitted?

Most submissions are assigned accession numbers within two working days of their arrival at GenBank. This time-frame may vary slightly depending on the volume of incoming submissions at the time you submit.

If you have not annotated your sequence or provided all the necessary information in your submission, you will be contacted for this information prior to the assignment of an accession number.

Note: the receipt of an accession number does not mean that your GenBank submission is processed and available online. See the answer to "How long does it take to process my submission?"for information on GenBank submission processing.

If you have not heard from us within two working days and wish to inquire about the status of your submission, do the following:

If you submitted via BankIt:

Email a message to gb-admin@ncbi.nlm.nih.gov, asking us to check the status of your submission. State in your message the email address that you used in your submission to GenBank, the BankIt ID, and the date you completed the BankIt submission process.

If you submitted via Sequin:

Email a message to gb-admin@ncbi.nlm.nih.gov, asking us to check the status of your submission. State in your message the email address that you used to submit to GenBank, the sequin file name, and the date you emailed the file.

How long does it take to process my submission? You sent my accession number, but I can't find my sequence in GenBank.

Submissions are not automatically deposited into the GenBank database after being assigned their accession numbers.

Your sequences will first be examined and processed individually by the GenBank annotation staff members to determine if they contain errors or problems. When your record is processed, we will contact you if we require additional information. **When your record is complete**, a final copy of your GenBank record will be sent to you, and the record will be made publicly available.

Acknowledgement of Submission

When I submit a sequence using BankIt, how will I know that GenBank actually got my submission?

Each time you send a submission to GenBank via BankIt, an automatic reply is generated and sent to the email address used in your submission. This automatic reply states that you will be hearing from the GenBank submissions staff within two working days.

If has been two working days since you completed the BankIt submissions process, and you haven't yet received a response, do the following:

Email a message to gb-admin@ncbi.nlm.nih.gov, asking us to check the status of your submission. Be sure to state in your message the email address that you used in your submission to GenBank, the BankIt ID, and the date you completed the BankIt submission process.

I emailed my Sequin submission to GenBank, but have not received an acknowledgement that my submission was ever received.

When we receive a new Sequin submission, an automatic reply is generated and sent to the email address used in your submission. This automatic reply states that you will be hearing from the GenBank submissions staff within two working days.

Make certain that you have actually emailed the .sqn submission files generated by the Sequin program to gb-sub@ncbi.nlm.nih.gov as instructed in the dialog box at the end of the submission. **Remember: Sequin does not automatically transmit the completed file for you at the end of the Sequin process**.

If it has been two working days since you emailed your .sqn files to GenBank's submissions staff, and you haven't yet received a response, do the following:

Email a message to gb-admin@ncbi.nlm.nih.gov, asking us to check the status of your submission. **Be sure to state in your message the email address that you used to submit to GenBank, the sequin file name, and the date you emailed the file**.

Accession Numbers

I've finished the Sequin process and sent in my Sequin submission, but I've not received my accession numbers.

Make certain that you have actually emailed the .sqn submission files generated by the Sequin program to gb-sub@ncbi.nlm.nih.gov as instructed in the dialog box at the end of the submission. **Remember: Sequin does not automatically transmit the completed file to GenBank for you at the end of the Sequin process**.

If it has been two working days since you emailed your .sqn files to GenBank's submissions staff, and you have not received accession numbers, you can:

- Email a message to gb-admin@ncbi.nlm.nih.gov, asking us to check the status of your submission. State in your message the email address that you used to submit to GenBank, the sequin file name, and the date you emailed the file.

Or

- **Resend your completed .sqn file togb-sub@ncbi.nlm.nih.gov, and send a separate confirmation email togb-admin@ncbi.nlm.nih.gov** indicating that you have emailed new submissions to GenBank. Include in this confirmation the email address from which you sent your submissions.

Changing a File or Record after Submission (Submission Updates)

Created: April 6, 2011; Updated: November 3, 2014.

Is it possible to make changes in a file after an accession number has been assigned?

Yes, you can submit an update to your GenBank submission after submitting, but remember:

Regardless of whether you are using Sequin or BankIt to submit your update, you must format the update as specified in the "Updating Information in GenBank Records"page.

Changing (Updating) a Record Using Sequin

How do I update my existing GenBank record using Sequin?

The GenBank Update page lists the proper formats for updating different kinds of data within your record. In order for your updates to be processed, they must be submitted in one of these formats.

Note: You can only use Sequin to update your record if:

1. **If we have assigned an accession number, but have not yet processed the record**:

   You can send us a new Sequin (.sqn) file as long as the sequence identifiers (SeqIDs) are exactly the same as those used in the original Sequin file.
2. **If the record is publicly available**:

   You must useNetwork Aware Sequinto download the existing record and edit it. Once you make the changes in Network aware Sequin, email the Sequin (.sqn) file containing the updated version to: gb-admin@ncbi.nlm.nih.gov.

   If your sequence has been assigned accession numbers, but is being held pending publication you should not use Sequin to update your record. See one of the methods on the GenBank Update page.

When I update a record using Network Aware Sequin, do I use the original submission file that I submitted to create the update?

Do not use your original Sequin submission file for a Network Aware update — **you must start your Network Aware Sequin update using the most recent version of the Sequin file for your record that GenBank has**, rather than the file that you originally submitted, since we will have changed the file during processing, and we need the changes we made in place for your update. You can download the latest file from the public database using Network Aware Sequin.

I have heard that Network Aware Sequin is the best alternative for creating a complex update to a record originally submitted using Sequin. What if I don't have the capability to use Network Aware sequin?

If you do not have the capability for Network Aware Sequin, there are two alternatives for sending your update information to us:

1 Email us at gb-admin@ncbi.nlm.nih.gov, and request a tab-delimited 5-column Feature table with your record's current annotation on it from us. You can then edit the feature table we send to you with the updated information and return it to us at gb-admin@ncbi.nlm.nih.gov.

OR

2. You can also send us the revised features in a spread sheet as delineated on the GenBank Update page.

Update Not from Original Submitter

The head of my lab submitted a sequence to GenBank some time ago, and wants me to update it. Can I submit an annotation update for her GenBank record even though I wasn't listed as a submitter on the original record?

Submitting authors of a GenBank record maintain editorial control of any record they submit. **Since you are not listed as an author on the original submission, GenBank will not be able to complete any update you submit unless** you contact one of the original submitters (the head of your lab or another member of the group listed as a submitter on the record) and have them contact GenBank at gb-admin@ncbi.nlm.nih.gov to give us permission to complete the update you submit.

Submitting Sequence Data when a GenBank Record for the Sequence Already Exists

I want to submit data that I sequenced, but have found that the same sequence was submitted to GenBank in 2001. Can I still submit my data? Will my sequence data replace the existing record?

GenBank is a redundant, non-curated database, therefore we accept the same sequence from different submitters. So even if your sequence is identical to an existing GenBank sequence record, you may submit your data as a new record to GenBank. It will have its own accession number and will be distinct from the sequence record that already exists.

If I submitted a gene sequence back in 2005, but recently was able to generate additional sequence data for it, do I submit the original sequence data and the newly generated sequence together as a new submission?

Do not send the original data and the new sequence data together as a new submission – send them together as an update.

Submitting Annotation Data for an Existing GenBank Record you did not Sequence

Although I'm not listed as an original submitter on an existing GenBank record, I'd like to submit additional annotation for it. How do I do this?

The original submitters of a sequence in GenBank maintain editorial control of any sequence they submit, and if you are not listed as an author on the original submission, we cannot complete any update you submit.

If you have new annotation for an existing GenBank record, visit GenBank's Third Party Sequence Annotation database (TPA). The TPA is a way to capture experimental or inferential annotation data for sequences submitted by another user.

There are specific criteria that your data need to meet for inclusion in the TPA database, so read through the information on the TPA web site carefully to verify that your data meets these criteria before submitting.

Note: all annotations submitted to TPA must be accompanied by experimental evidence (direct or indirect), showing the existence of the annotation submitted.

Look for examples of experimental data that can be sent to the TPA database on the TPA:experimental page and examples of inferential data that can be sent to the TPA database on the TPA:inferential page. A list of data that should not be sent to TPA is located at the bottom of the TPA home page (scroll to bottom of page).

A User's Guide to BankIt

Michael Fetchko and Adrienne Kitts

The Design of this User's Guide

Michael Fetchko and Adrienne Kitts

This User's Guide is designed to be a practical, plain language guide for the beginner— or for anyone who wants basic instructions for any of the steps you must perform when you submit using BankIt.

Each major section in this User's Guide represents one of the pages in the BankIt submission tool. In each section you will find a summary of purpose for the BankIt page and step-by-step instructions for the more detailed parts of the page. These instructions include screen captures that will help you relate the instructions you see in this guide to the BankIt submission tool pages.

At the end of each section of the User's Guide, there is a discussion of some of the common mistakes people make when filling out that particular page of BankIt tool, and how to fix them.

What is BankIt?

Michael Fetchko and Adrienne Kitts

BankIt is a web-based sequence submission tool. Use it to submit to GenBank if you want to submit:

- A single sequence
- A few unrelated sequences or a few sequences with different features and/or source information
- A large set of sequences with a small number of the same features/source information
- A small batch of sequences with a small number of features or source information

To complete a BankIt sequence submission, you will be prompted by the BankIt program to provide:

- Contact Information for the person performing the submission
- Reference information
  - The name(s) of the sequence authors
  - The title of the intended or published paper describing the sequence
- Nucleotide sequence and general information about the sequence
- Submission Category (was the submission directly sequenced or created from existing primary data)
- Source Information (information about the sample or life form from which the sequence was isolated)
- PCR Primers (this information is optional)
- Feature Information (information about each part of the sequence that has a specific function: e.g. exon, intron, gene, coding region, functional RNA, etc.)

Once you have entered all your submission information, the last page of the BankIt tool will display the finished flatfile made from the information you provided. You will also have a chance at that point to fix any mistakes or add additional data.

BankIt's Multi Page Tab Design

Because you will need to provide different types of information with the sequence(s) you are submitting, the BankIt submission tool uses separate pages that accept different types of information for a submission. Each page is labeled with a tab, and all tabs are displayed at the top of the main BankIt page.

Tab Labels Indicate Required Information or Action(s)

Each tab shows the type of information you will need to provide, or an action you need to perform.

For example, the tab marked "Nucleotide" denotes a page where you will provide your sequence data and any additional information that describes the type of molecule you are submitting. The "Review and Correct" tab, allows you to review the flatfile made from the information you submitted in the previous pages, and gives you the chance to make corrections to the data before you finish the submission.

Tab Links will Activate as you Complete Form Pages

When you start your submission, all of the tabs are grey and are not active. A tab will turn black when you begin to fill in the page, and then will turn color and become active as a link Figure 1 once you have successfully filled out the page and clicked the "Continue" button (to move to the next page). These tab links allow you to go back to any of the pages of the tool once you have completed them, but prevent you from filling in the pages out of order.

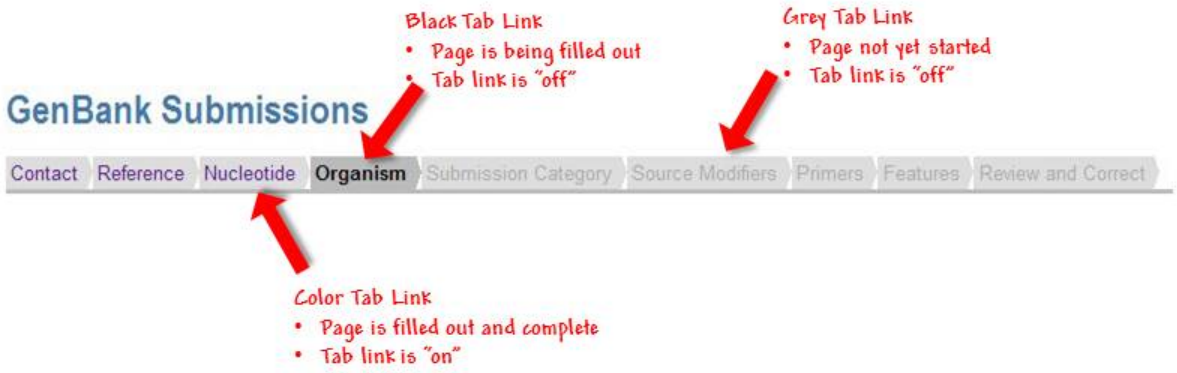

Figure 1: Page tabs on the BankIt submission form. Figure text shows page and tab link status.

Navigating through the BankIt Form: Tabs vs. your Browser's Back Arrow

It is essential that you navigate through the BankIt tool using the links on the BankIt page tabs. Do not navigate through BankIt using your browser's Back or Forward arrows once you have started a submission. If you navigate to a BankIt page using your browser's Back or Forward arrows, you may lose the information that you entered for your submission.

Messages from the BankIt Program

If you leave out necessary data or make a mistake as you fill in a BankIt page, once you finish the page and click the "Continue" button to go to the next page, you will be stopped by BankIt.

Instead of moving to the next page, you will be taken to the top of the page you just filled out where a message (Error, Warning, or Note) will be displayed in a light blue box. This message will either give you instructions for an action you need to perform, or, if there is a mistake or items missing from the data you put in the page, give information to help you figure out the problem and fix it. Error messages must be resolved before you can move on to the next BankIt page.

Once you fix any problem and click the “Continue” button, BankIt will allow you to move to the next page.

Returning to a Completed Page before Fixing an Error in the Current Page

If you see an error message on a page you just finished filling out and then click on a tab for an earlier completed page before you fix the error, your data will **not** be saved —you will lose all the data you entered in the current page.

No warning dialogue box will pop up. When you click on the tab to a completed page, you will go to that page, but when you go back to the page that had the error warning, the entire page will be empty of anything you entered there. You will have to re-enter the information for that page again.

As long as you fix any errors in your current page and click “Continue” before clicking a tab to an earlier completed page, all of the data you enter into the current page will be saved.

Questions about using BankIt

If you have questions about using BankIt while you work on a submission, contact GenBank User Services at: info@ncbi.nlm.nih.gov . Please include your BankIt submission number (shown at the top of each BankIt page) in any emails you send.

The "Contact" Page

Michael Fetchko and Adrienne Kitts

Purpose

The purpose of the BankIt submission tool "Contact" page is to get contact information for the person doing the submission. This person does not have to be one of the data authors, nor does the person have to be an author of the publication that describes the data. The submitter can be a member of the laboratory that produced the data, or anyone that is able to get the information needed to answer questions we may have about the data submitted.

The information you supply in this section will be kept confidential and will only be used to contact you if we have questions about your submission.

New Submitters

Once you create a Primary Data Archive (PDA) account and use it to log into BankIt, you will be taken to the BankIt site. Once there, click the "New Submission" button to go to the first page of the BankIt submission tool called the "Contact" page, which will have the name and the email address you gave when you created your PDA account.

In order to use the rest of the BankIt tool, you must first provide the information requested on the "Contact" page, which includes:

- First (given) and last (family) names
- The name of your department
- The name of your institution
- The full address of your institution
  Note: "State/Province is optional for some countries.
- Telephone number
- Fax number
- Email
- Alternate email

Once you have provided the information, click the "Continue" button at the bottom of the page. If the "Contact" page has been filled out correctly, you will go to the next page. Also, once you click the "Continue" button, the contact information you provided will be saved so that the next time the Contact page will already be completed. If any information is missing when you click the "Continue" button, BankIt will ask you to supply the information before you are allowed to continue with your submission.

Users with an existing BankIt Account

If you have already submitted using BankIt, the next time you login to BankIt and click the "New Submission" button, you will go to the BankIt "Contact" page, where you will find that the contact information that you provided before was saved. Review your contact information and update it if necessary before you continue your new submission.

Alternate Email Addresses

Although an alternate email address is not required for the "Contact" page, it is helpful if you enter a personal email address in the "Alternate Email" field. The alternate email address gives us a way to contact you about your

submission if you should change your institute or organization. This is especially important if you set a release date for your sequence that is a year or more beyond the date of your submission.

Common Mistakes Made While Filling out the Contact Page

You may see an error or warning message on the Contact page if you do not provide us all of the contact information we ask for on the page. You need to provide:

- A given (first) name
- A family (last) name
- A complete mailing address
- A telephone number
- An email address.
- A fax number, if you have one available.

The "Reference" Page

Michael Fetchko and Adrienne Kitts

Purpose

The purpose of the BankIt submission tool "Reference" page is to collect information about the project or paper for the data you are submitting. This includes:

- The names of the people who contributed to the sequencing of the nucleic acid sequence you intend to submit
- A draft title for the journal article that discusses this sequence (if it is unpublished or you are in the middle of writing)
- A title for the journal article that discusses this sequence (if the article is in press or is already published)
- The names of the authors of the journal article

The "Sequence Authors" Section

In this section provide the first name, middle initial(s) and last name of the people who contributed in some way to the actual sequencing of the data being submitted (see figure 2). This section is important and must be filled out carefully according to the instructions provided in Box 1.

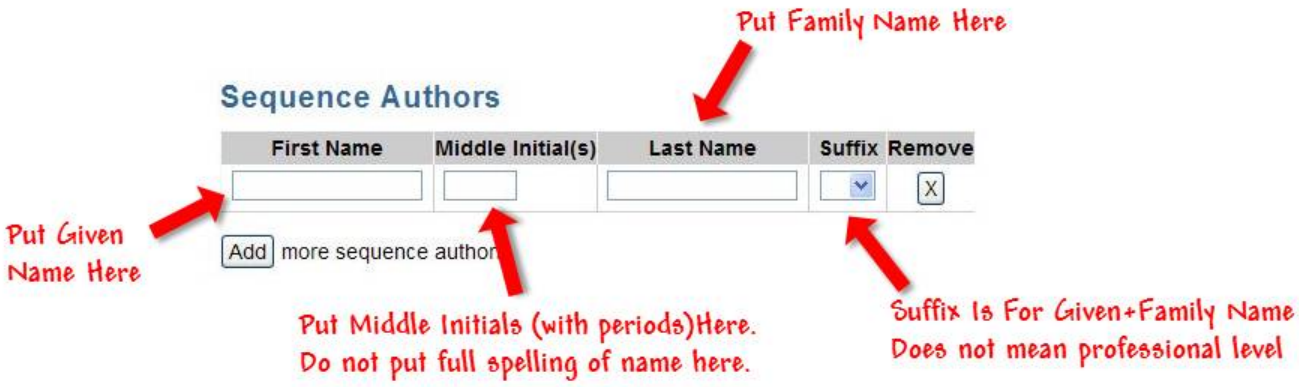

Figure 2: The default "Sequence Authors" section of the BankIt form "Reference" page.

Box 1: How to Enter Sequence Author Names:

Starting from the left side of the page:

1. **Enter your given name in the box marked "First Name".**
   (In English, for the name John Smith, the given name is John. In Chinese, for the name Yao Ming, the given name is Ming. In Japanese, for the name Yamada Hanako, the given name is Hanako).
2. **Enter your family name in the box marked "Last Name".**
   (In English, for the name John Smith, the family name is Smith. In Chinese, for the name Yao Ming, the family name is Yao. In Japanese, for the name Yamada Hanako, the family name is Yamada).
3. **Enter the initial(s) of your middle name(s) in the box marked "Middle Initial(s)".**
   Even if you like using the full spelling of your middle name, and an initial for your given (first) name: "A. Jane Smith", you must still enter the full spelling of your given (first) name, and an initial for your middle name: "Abigail J. Smith" when you use this form.

Box 1 continues on next page...

Box 1 continued from previous page.

4. **Enter your Family name in the box marked "Last Name".**
   Be sure to provide the complete spelling of your family name, even if it has more than one word in it (e.g. Bowes-Lyon or Vaughn Williams)
5. **Enter the suffix to your given+family name in the box marked "Suffix".**
   A suffix does not refer to your seniority in your laboratory or institute— it refers to a personal name (e.g. John Smith, Jr. or John Smith III or John Smith IV).

Adding more Authors

Initially, the 'Sequence Authors" section of the Reference page has space for only one sequence author. If you need to add more sequence authors, click the "Add" button, located just below the space for the sequence author. Each time you click the "Add" button, another space for another sequence author will appear (see figure 3). You can click the "Add" button as many times as you have sequence authors.

Deleting an Author

Click the "X" button marked "Remove" located to the far right of each name entry (see figure 3) to remove an author from the list of sequence authors. It does not matter if the name you wish to remove is in the middle of the list, the remaining names will stay in the same order.

The "Reference Information" Section

In this section provide information about the paper that discusses (or intends to discuss) the data you are submitting. Even if the paper is not published, or has not even been written yet, we still want you to provide some information about it: a title or basic description of the study, publication information (if any), and authors. See figure 4 for the preset design of the Reference information section.

Filling in the "Publication Status" Subsection

Here, provide us with information about the current stage of publication of the paper and a list of its authors.

The "Unpublished" Button

Use this button if you have not submitted anything for publication yet (i.e. if you are in the middle of writing or you haven't written anything at all) or if you have submitted a paper, but it has not yet been accepted.

If "unpublished" is the status of your project you must provide a reference title:

- **If you are in the middle of writing, give the title of your unfinished document** — even if this title changes later on, the submission can be updated with the final title. Thinking of a title will help you describe the sequence you are submitting, and why you sequenced it.
- **If you haven't written anything yet, you still must provide a reference title** —it should be a short description of the sequence you are submitting and why you sequenced it (a short description of your proposed paper that includes the sequence you are submitting).
- **Do not use Reference Titles like "Direct Submission" or "Not Published".**

The "In-Press" Button

If the document you have written has been accepted for publication, but it hasn't been released yet, select the "In-Press" button. When you select this button, new boxes will appear below the "Reference Title" box (see figure 5):

Figure 3: The "Sequence Authors" section of the BankIt form "Reference" page after the "Add" button has been clicked 3 times.

1. Fill in the Reference Title box with the title of your paper.
2. Fill in the new boxes that appear below the Reference Title box with general publication information like the journal title, the volume, issue, pages and year of publication (the publication year must be equal to or greater than the current year).

The "Published" Button

If your paper has been published, select the "Published" button. When you select this button, new boxes will appear below the "Reference Title" text box. Instructions for filling in these boxes are in Box 2.

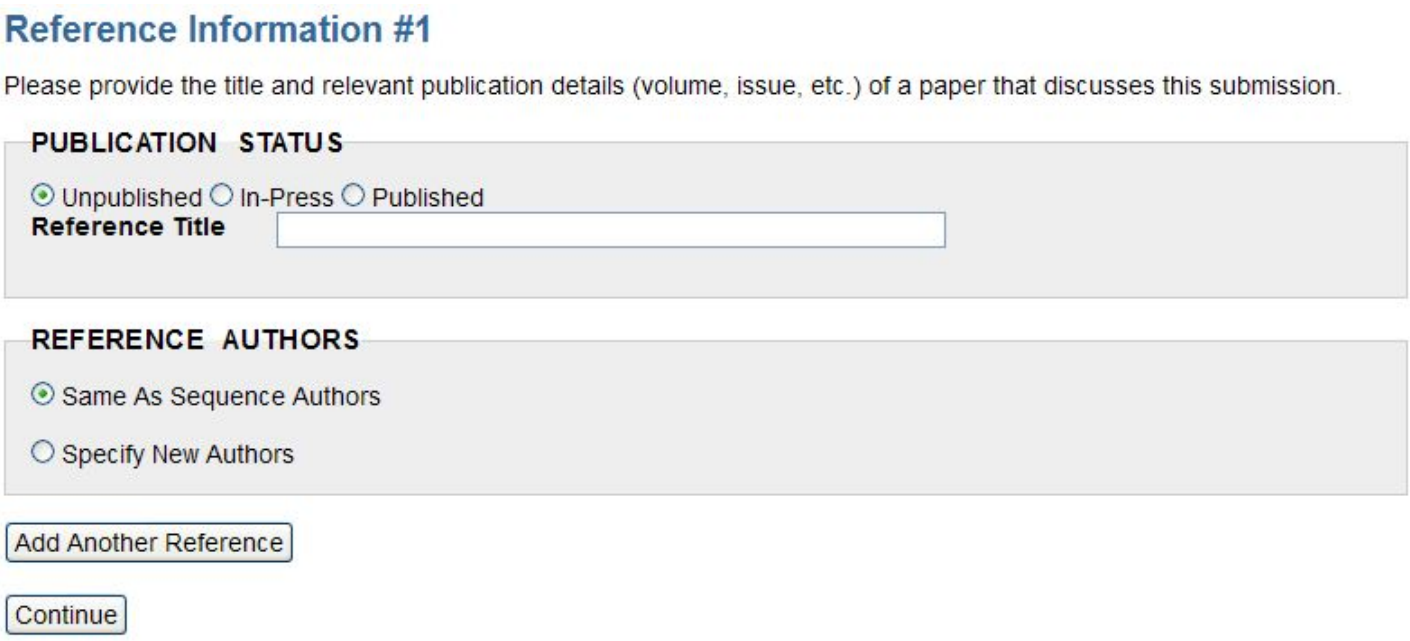

Figure 4: The preset design of the Reference Information section before status or author selections have been made.

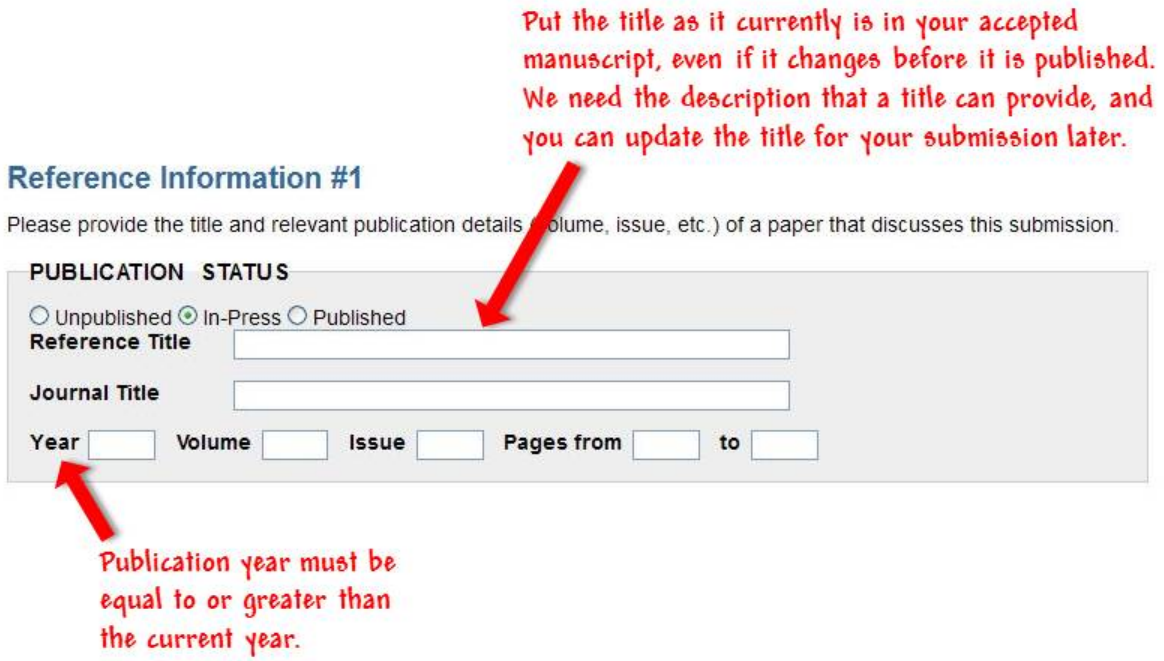

Figure 5: The Reference Information section after the "In-Press" button has been clicked. Figure text provides hints for filling out the "In-Press" section.

Box 2: Filling out the "Publication Status" Subsection for a Published Paper.

If you have the PubMed ID (PMID) for your paper:

1. **Fill in the bottom box with the PMID of your paper** (see Figure 6).
2. **Click the "Continue" button at the bottom of the page**.
3. A reference in a blue box will appear if the PMID you entered matches an actual PMID number. **Check the reference carefully to be sure it is the right one** (see Figure 7).
4. **If the PMID you entered gives you the correct reference**, click the "yes" button just below the reference, and click the "Continue" button to go to the next page of the BankIt form. (see Figure 7)
5. **If the PMID you entered does not give you the correct reference**, click the "no" button just below the reference, and the grey "Publication Status" section will reappear below the blue box of reference data. (see Figure 8)
6. Fill in the correct PMID and click the "Continue" button at the bottom of the page.
7. Repeat steps 3 through 4 or 5

Box 2 continues on next page...

Box 2 continued from previous page.

8. If you still did not get the correct reference, instead of using the PMID, you may want to fill in the general publication information that is asked for in the Publication Status box (Reference Title, Journal Title, year, pages).
9. **Be sure to remove the incorrect PMID from the box after filling in the publication information asked for in the "Publication Status" box**. If you do not remove the incorrect PMID from the box, the incorrect reference for that PMID will keep coming back every time you click the "Continue" button even if you have filled in the information asked for in the "Publication Status" box.
10. Once all of your references are correctly entered, click the "Continue" button at the bottom of the page to go to the next page of the BankIt form.

If you do not have the PubMed ID (PMID) for your paper:

1. Fill in the Reference Title box with the title of your paper.
2. Fill in the boxes that ask for the general publication information of your paper.
3. Proceed to the Reference Authors Subsection to provide author information for your publication.

Filling in the "Reference Authors" Subsection

In this section, provide us the first name, middle initial(s) and last name of the people who are (or will be) authors of the publication (or potential publication) that discusses the data being submitted.

The "Same as Sequence Authors" Button

If the authors of the publication (or potential publication) are (or will be) the same as the sequence authors, use the preset "Same as Reference Authors" button.

The "Specify New Authors" Button

If the authors of the publication are (or will be) different than the sequence authors, click the "Specify New Authors" button. When you click this button, you will get a series of new boxes you must fill in for the authors of the publication. Instructions for entering Reference Author names are available in Box 3.

Box 3: How to Enter Reference Author Names:

Start from the left side of the page:

1. **Enter your given name in the box marked "First Name".**
   (In English, for the name John Smith, the given name is John. In Chinese, for the name Yao Ming the given name is Ming. In Japanese, for the name Yamada Hanako, the given name is Hanako).
2. **Enter your family name in the box marked "Last Name".**
   (In English, for the name John Smith, the family name is Smith. In Chinese, for the name Yao Ming the family name is Yao. In Japanese, for the name Yamada Hanako, the family name is Yamada).
3. **Enter the initial(s) of your middle name(s) in the box marked "Middle Initial(s)".**
   Even if you like using the full spelling of your middle name, and an initial for your given (first) name: "A. Jane Smith", you must enter the full spelling of your given (first) name, and an initial for your middle name: "Abigail J. Smith" when you use this form.

Box 3 continues on next page...

Box 3 continued from previous page.

4. **Enter your Family name in the box marked "Last Name".**
   Be sure to provide the complete spelling of your family name, even if it has more than one word in it (e.g. Bowes-Lyon or Vaughn Williams)
5. **Enter the suffix to your given name in the box marked "Suffix".**
   A suffix does not refer to your seniority in your laboratory or institute— it refers to a personal name (e.g. John Smith, Jr., John Smith III, or John Smith IV).

Adding more Authors

Initially, the 'Reference Authors" section has space for only one reference author. If you need to add more reference authors, click the "Add" button, located just below the space for the reference author. Each time you click the "Add" button, another space for another reference author will appear. You can click the "Add" button as many times as you have reference authors.

Deleting an Author

In order to remove an author from the list of reference authors, click the "X" button marked "Remove" located to the far right of each name entry. It does not matter if the name you wish to remove is in the middle of the list, the remaining names will stay in the same order.

Add another Reference for your Submission

If the sequence you are submitting to us has (or will have) more than one paper connected to it, click the "Add Another Reference" button once you finish filling out the information about the first reference. This button is located at the bottom of the "Reference" page of the BankIt form.

Once you click this button, another "Reference Information" subsection will appear below the first one you filled out. Fill in this new reference subsection the same way you did the first using the information for the additional reference.

Common Mistakes Made While Filling Out the "Reference" Page

- **Mistake: Entering the first and last names of the sequence author(s) in the wrong order:**

  Fix: Enter your given name in the "First Name" box, and enter your family name in the "Last Name" Box.
- **Mistake:Entering a full name in the box asking for middle initials**

  Fix: Even if you use the full spelling of your middle name, and an initial for your first name, you still must put a full given name in the "First Name" box, and only your middle name(s) initial(s) in the "Middle Initial(s)" box.

  For example, if your full name is Jane Abigail Smith, but you like to be known as J. Abigail Smith, you still must put "Jane" in the "First Name" box, and "A." (no quotes) in the "Middle Initial(s)" box.
- **Mistake: Using the Suffix Box to indicate your professional level**

  Fix: A suffix should be used only if your personal name requires it. A suffix does not refer to your seniority in your laboratory or institute — it refers to a personal name that has been passed down within a single family over a number of generations.

For example, John Andrew Smith is the first individual in the Smith family tree to have that name, while John Andrew Smith IV is the fourth individual in the Smith family tree to have that name.

- **Mistake:Putting only one part of a multi word last (family) name in the "Last Name" box.**

  Fix: Always put the complete spelling of your family name in the "Last Name" box, even if it has more than one word in it (e.g. Bowes-Lyon or Vaughn Williams).
- **Mistake: Not providing the title of an unpublished work in the "Reference Title" box**

  Fix: Even if you haven't written anything yet, you still must provide a reference title: a short description of the sequence you are submitting and why you sequenced it (a short description of the paper that includes the sequence you are submitting).

  If you are in the middle of writing, give the title of your unfinished document — even if this title changes later, the submission can be updated with the final title. Thinking of a title will help you describe the sequence you are submitting, and why you sequenced it.

  Do NOT use reference titles like "Direct Submission" or "Not Published". These names don't describe your sequence or its importance.

Other Mistakes that are commonly made on the reference page include:

- Putting numbers in a box where text should go
- Putting letters in a box where numbers should go
- Entering an invalid PMID

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.

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.

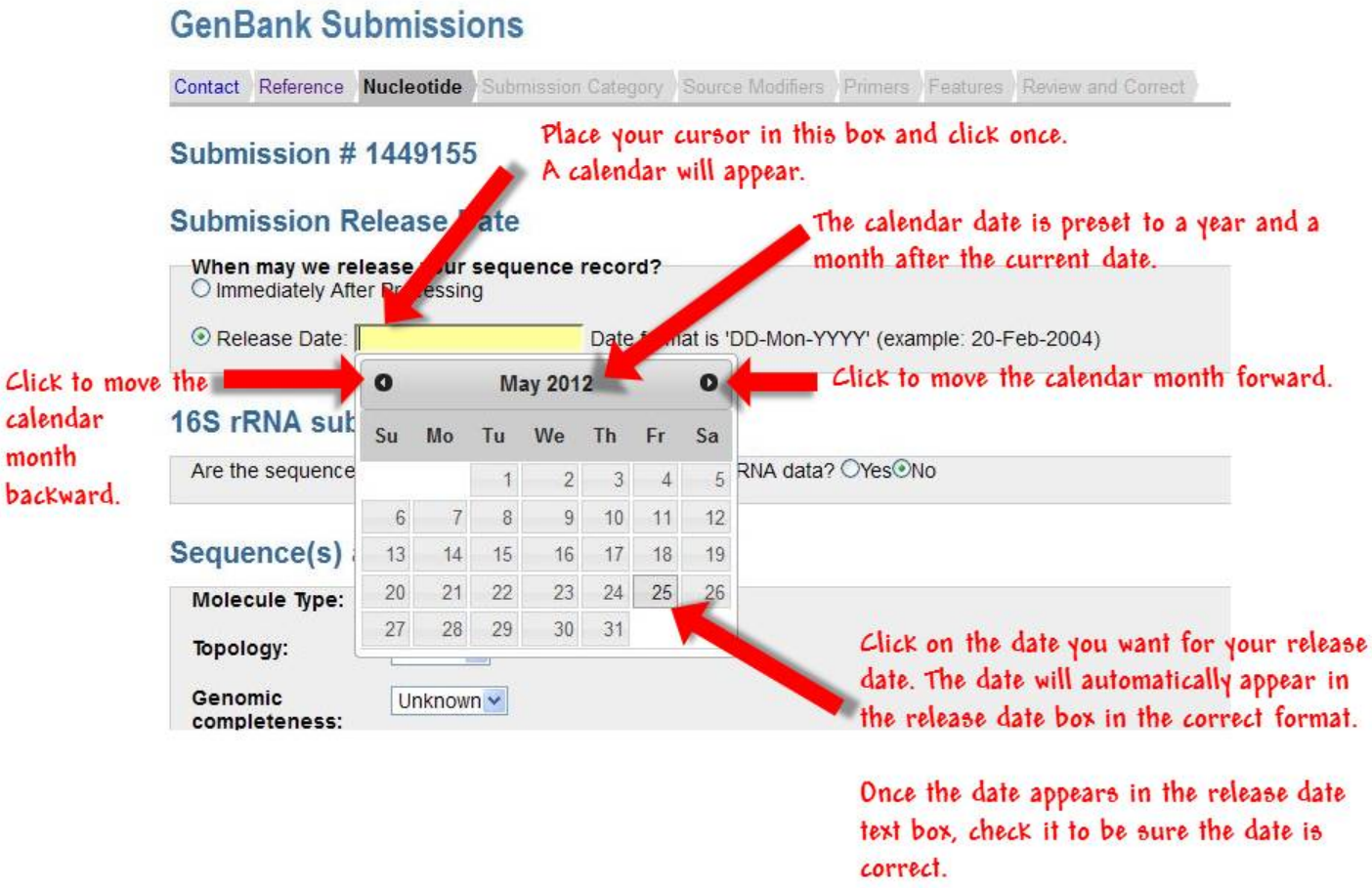

Figure 9: The Submission Release Date section of the Nucleotide page showing the calendar that appears once you place your cursor in the text box and click. Figure text provides hints for using the calendar.

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.

The "Organism" Page

Michael Fetchko and Adrienne Kitts

This page will appear after the "Nucleotide" page only if you made an error in the organism FASTA format or did not provide an organism in the FASTA definition line at all. If you correctly entered an organism in the definition line for each sequence you are submitting in the "Nucleotide" page, the "Organism" page will not appear, and you will go directly to the next page of the form.

If you have a single sequence, or multiple sequences from the same organism:

1. Begin entering the organism name in the text box to the right of "Organism name":
2. Once you begin entering the organism name in the text box provided, BankIt will provide a list of species names based on the text you enter (Figure 10).
3. Select the appropriate organism name for your submission. If you do not see the appropriate organism listed, type the correct and complete organism name into the text box.
4. Click the "Continue" button.
5. You will see the organism name appear with its sequence ID in a static table below "Continue" button. Check to be sure the organism you selected or typed in is correct.
6. Click on the "Continue" button again to go to the next page of the submission form.

If you have multiple sequences with multiple names:

1. Select an organism file from your local computer using the "Browse" button to the right of the words "Upload File". See the BankIt Help documentation for instructions for creating an organism file.
2. Click the "Continue" button.
3. You will see the organism names and sequence IDs from your organism table appear in a static table below "Continue" button. Check to be sure the organism names and sequence IDs are correct.
4. Click on the "Continue" button again to go to the next page of the submission form.

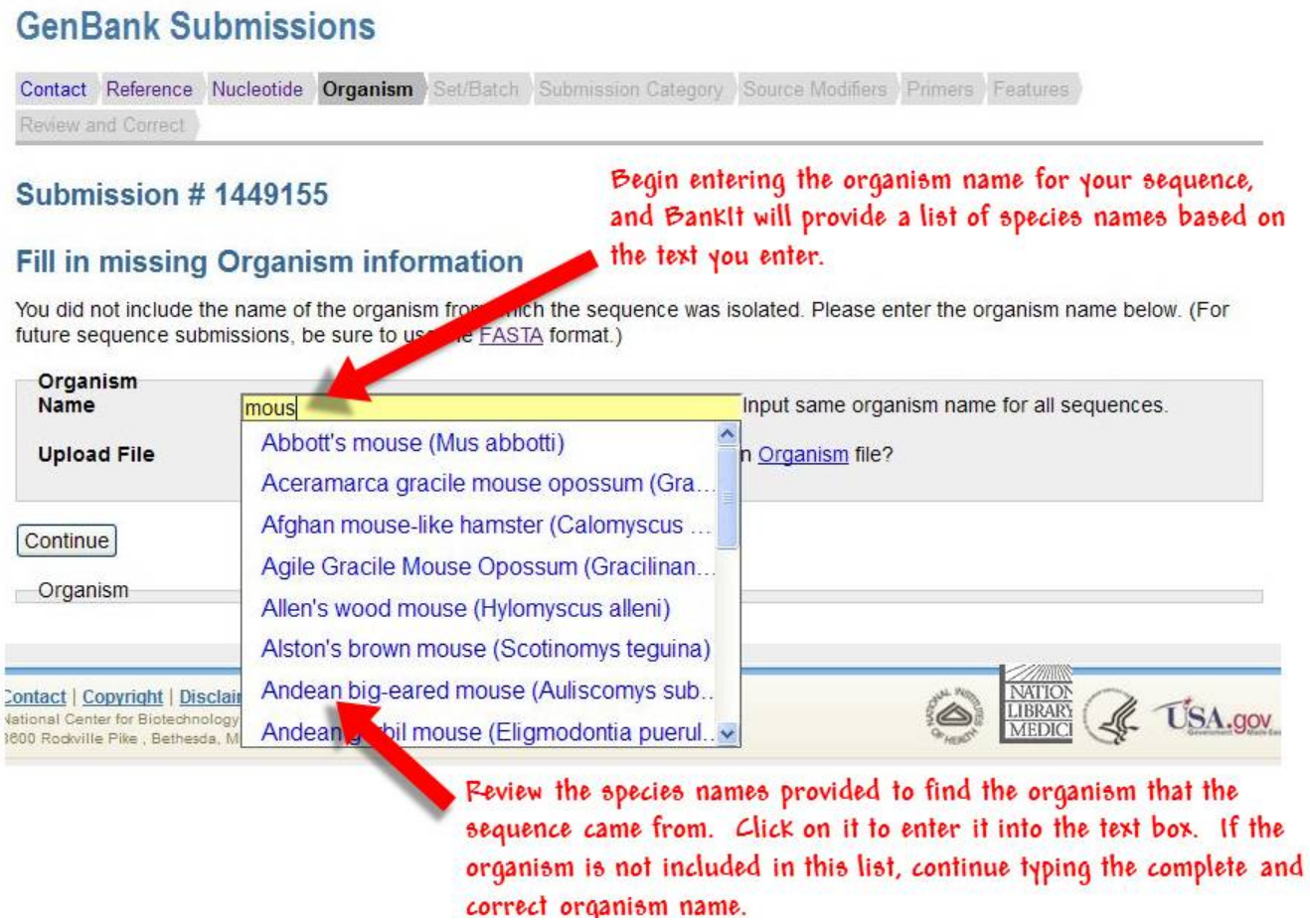

Figure 10: The Organism page showing a species drop down list released as you enter text in "Organism Name" text box. Figure text gives information about the drop-down list of species names.

The "Set/Batch" Page

Michael Fetchko and Adrienne Kitts

This page will appear after the "Nucleotide" page if you entered more than one sequence. If you entered a single sequence in the "Nucleotide" page, the "Set/Batch" page will not appear, and you will go directly to the next page of the form.

Select one of the set types listed and defined in the Set/Batch page to identify what kind of set you are submitting, or select the "Batch" option. Select the "Batch option if the nucleotide sequences you are submitting are not of the same gene, but are related because they may be discussed in the same publication or are from the same organism.

The "Submission Category" Page

Michael Fetchko and Adrienne Kitts

Purpose

The BankIt submission tool "Submission Category" page is where you confirm that you actually sequenced the data you are submitting. If you did not sequence the data you are submitting but want to create Third Party Annotation (TPA) submission, you will be asked to provide:

- A description of the biological experiments or other work used as evidence for the new annotation in your TPA submission
- The GenBank Accession number(s) of the primary sequence(s) from which your TPA submission was derived.

Original Submissions

The pre-set answer of "Original" indicates that the sequence is an original submission that was sequenced directly by the submitter. If you leave the "Original" button selected and click the "Continue" button at the bottom of the page, you will proceed directly to the next page of the BankIt form.

What is an Original Submission?

All of the following are considered original submissions:

- Data sequenced directly by the submitter

  Note: If your sequence is identical to an existing GenBank sequence record, your sequence is still considered original, and must be submitted as new data to GenBank. It will have its own accession number and will be distinct from the sequence record that already exists.
- Synthetic sequences
  - a whole cloning vector that you designed
  - codon-optimized genes/coding sequences for use in specific organisms
- Sequence amplified using PCR primers derived from other sequences

Third Party Annotation (TPA)

If you click the "Third Party Annotation" button, a submission form for the Third Party Annotation (TPA) sequence database will appear. A TPA submission adds new feature annotation for primary sequences (i. e. sequences available in GenBank that already have GenBank Accession numbers) that is supplied by experimental or inferential evidence.

A TPA sequence must be built from primary sequence data available in GenBank and identified by a GenBank accession number.

There are specific standards that your data must meet to be included in the TPA database, so read through the information on the TPA web site carefully to be sure that your data meets these standards before you submit.

Evidence

Provide text describing the evidence that supports your new annotation to the primary sequence.

This evidence can be:

- Experimental: Your annotation is supported by wet lab evidence published in a peer-reviewed scientific journal

  OR
- Inferential: Your annotation is inferred from other work you did, but this work was not by direct experimentation. The supporting information for your inferred annotation must be published in a peer-reviewed scientific journal.

GenBank Accessions

Provide a file of all the GenBank Accession numbers for the primary sequences that you used to build or derive your TPA sequence. The online BankIt Help documentation includes detailed instructions for creating a TPA file of primary accession numbers.

Primary sequence data includes Whole Genome Shotgun (WGS) data or Trace Archive data, but NOT Reference Sequence (RefSeq) data or data from the CON (Contig) division of GenBank (you will see the word "CON" in the locus line of these records), since RefSeq and CON data are not primary sequence data.

Common Mistakes Made While Filling Out the "Submission Category" Page

- **Mistake: Entering an incorrect GenBank Accession number**

  Fix: If you get an error message saying that one of your GenBank Accessions is invalid, **make sure that the accessions were typed correctly and that**:
  - **The accession is NOT for a** Reference Sequence (RefSeq) record

    RefSeq accession numbers can be distinguished from GenBank accessions by their format of 2 alphabetic characters followed by an underscore character ('_') and then a series of numbers. For example, a RefSeq mRNA accession is NM_123456.
  - **The accession is NOT for a**CON (Contig) division record (you will see the word "CON" in the locus line of these records).

    Con division records are already built from other primary sequence(s) and therefore cannot be cited as a primary sequence)

The "Source Modifiers" Page

Michael Fetchko and Adrienne Kitts

Purpose

The BankIt "Source Modifiers" page is the place where you will give us information about the organism(s) from which you obtained the nucleic acid sequence(s) you are submitting. The information we ask for in this page includes:

- The cellular location (if applicable) of the nucleic acid sequence
- Descriptive information about the organism
- Any information about how or where the organism was obtained
- Any additional descriptive information about the organism(s) that will serve to more specifically identify the nature of the organism(s).

Bacterial/Archaeal Sequences

This section will appear on the "Source Modifiers" page only if you identified the organism of the sequence(s) you are submitting as being either bacteria or archaea on the "Nucleotide" (or "Organism") page.

This section will help you determine the required source modifiers you need to submit. The required source modifiers depend on the method you used to obtain the sequence from the bacteria or archaea. When you select one of the listed methods, the required source modifiers that you need to provide will appear in a light blue box just below (Figure 11).

You must select one of the methods listed in order to proceed with you submission. If you do not select one of the methods, when you click the "Continue' button, you will get an error message and will not be able to go to the next page of the form.

In addition to the required source information for your bacterial/archaeal sequence(s), make sure to submit any other source information that you have about your bacterial/archaeal sequence(s).

Mouse and Rat Submissions

Provide the strain of the mouse or rat from which you obtained your sequence(s) if it is available. If you do not provide the strain, you will receive a warning asking you to do so when you click on the "Continue" button to go to the next page of the form.

Once you enter your strain information, click the "Continue" button. This will update the static table at the bottom of the page with your source modifier selections. Review the table to make sure your source modifiers are correct, and then click the "Continue" button again to move to the next page of the form.

Rice Submissions

Provide the rice cultivar from which you obtained your sequence(s) if it is available. If you do not provide the cultivar, you will receive a warning asking you to do so when you click on the "Continue" button to go to the next page of the form.

Once you enter your cultivar information, click the "Continue" button. This will update the static table at the bottom of the page with your source modifier selections. Review the table to make sure your source modifiers are correct, and then click the "Continue" button again to move to the next page of the form.

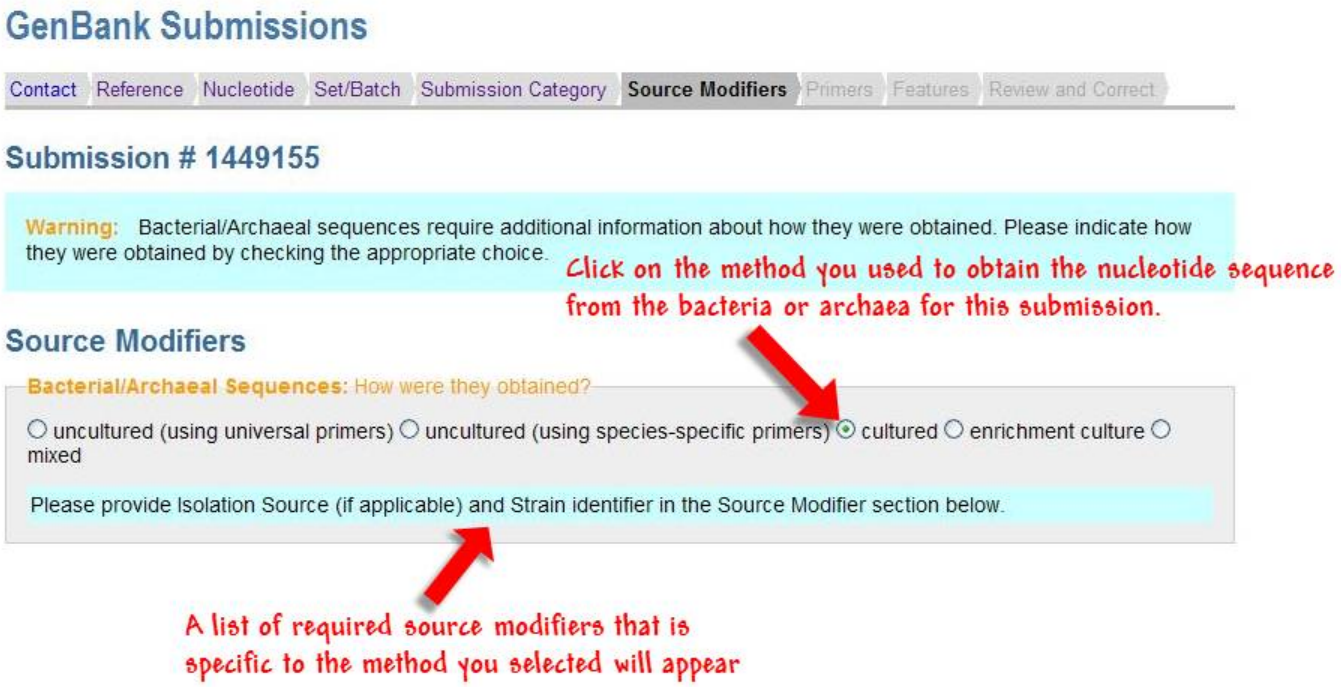

Figure 11: The Source Modifiers page when bacteria/archaea is submitted. This view shows the required source modifiers for the submission of bacteria/archaea sequences once a method is selected. Figure text gives information on how to get the list of required source modifiers.

Virus Submissions

You must provide the following information for virus submissions:

- A virus strain name or isolate name
- The country where the virus was isolated
  A country modifier must be from GenBank's approved list
- The date that the virus was collected
  A "collection_date" must be in the format DD-Mmm-YYYY.
  Examples: 12-Sep-2002 or Jun-1999 or 2010

If you do not provide one or more of these pieces of information, you will receive a warning message asking you to do so when you click the "Continue" button to go to the next page of the form.

If a country name is unrecognized or a collection date is provided in the wrong format, you will receive an error message when you click the "Continue" button to go to the next page of the form. The error message explains the

error and instructs you how to correct it. Once you correct the error, click the Continue button again to move to the next page.

Once you enter the requested virus information, click the "Continue" button. This will update the static table at the bottom of the page with your source modifier selections. Review the table to make sure your source modifiers are correct, and then click the "Continue" button again to move to the next page of the form.

The "Source Information" Section

Click the "Organelle/Location text box to release a drop-down menu of organelles and locations to choose from. (Figure 12). Select an organelle or location from the list only if it applies to your sequence.

Note:

- You do not have to select any of the listed organelles or locations if none of them applies to the sequence(s) you are submitting.
- The terms "nucleomorph" and "macronuclear" do NOT mean "nuclear". Do not choose either of these organelles unless they are appropriate to your specific organism

The "Source Modifiers" Section

Single Sequence Submissions

If you submitted a single sequence in the "Nucleotide" page of the form, when you get to the "source modifier" section of the "Source Modifiers" page, you will see a "source modifier" text box and a "value" text box.

Select source modifier(s) for your sequence by clicking in the "source modifier" text box to release a drop-down menu of source modifiers you can select from (Figure 13), and then enter the information that describes that source modifier in the "value" text box.

For example, if you select "collection date" as the source modifier, you would enter the date that the organism (from which you obtained the sequence) was collected in the "value" text box.

- If you are submitting a single sequence, you must enter the source modifiers for your sequence using the "source modifier" and "value" text boxes in this section of the page. You cannot upload a pre-made source modifier table.
- For each additional source modifier you have, click the "Add" button to generate additional "source modifier and "value" boxes you need to enter the information you have.

Once you have entered your source modifiers, click the "Continue" button. This will update the static table at the bottom of the page with your source modifier selections. Review the table to make sure your source modifiers are correct, and click the "Continue" button again to go to the next page of the form.

Multiple Sequence Submissions

If you submitted more than one sequence in the "Nucleotide" page of the form, when you get to the "source modifier" section of the "Source Modifiers" page, you will see two sub-sections: "Set multiple values for sequences" and "Set one value for all sequences" (Figure 14).

- **If the source modifier(s) you have are unique to each sequence you are submitting**, use the "Set multiple values for sequences" subsection and click on the "Upload source modifiers Table File" button. When you do, a "Browse" button will appear that you can use to select the source modifier table file from your computer (Figure 15).
  Once you select your table file, it will upload when you click the "Continue" button, and the static table at

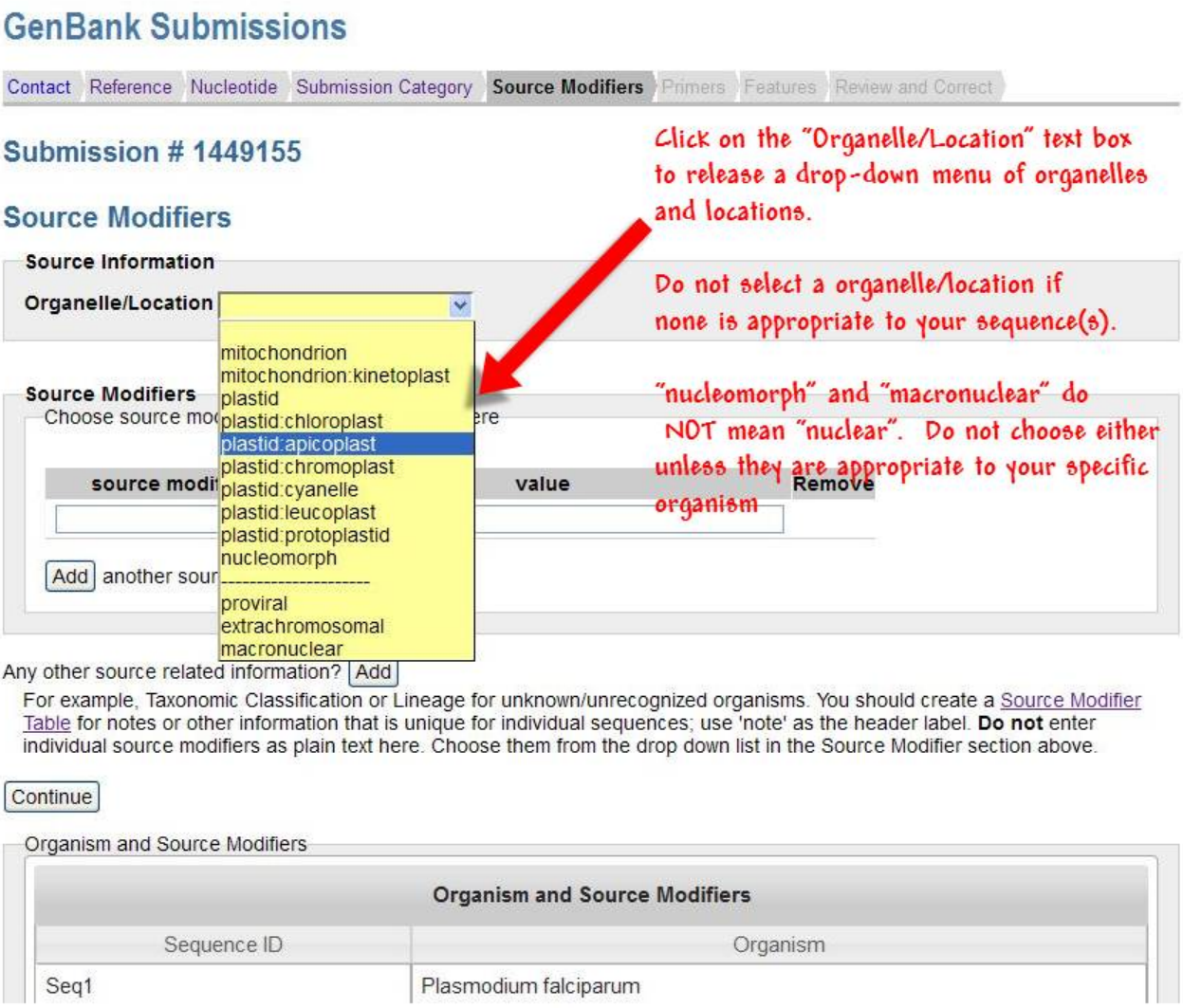

Figure 12: The "Source Information" section of the "Source Modifiers" page showing the drop-down list that appears when you click in the "Organelles/Locations" text box. Figure text gives hints for selecting the correct organelle/location.

the bottom of the page will update with your source modifier selections. Review the table to make sure your source modifiers are correct, and then click the "Continue" button again to move to the next page of the form.

If a modifier and/or value must be corrected, make the correction in your source table file and upload the corrected file, which will overwrite your existing selections.

For instructions on how create a source modifier table file, see the GenBank Submissions Quick Start, or use the BankIt Help documentation.

Note: If the source modifiers you have are unique to each sequence you are submitting you **must** submit the source modifiers for your sequences by uploading a source modifier table.

OR

- **If the same source modifier(s) apply to all the sequences you are submitting**, use the "Set one value for all sequences" subsection and click the "Choose source modifier" button. When you do, "source modifier"

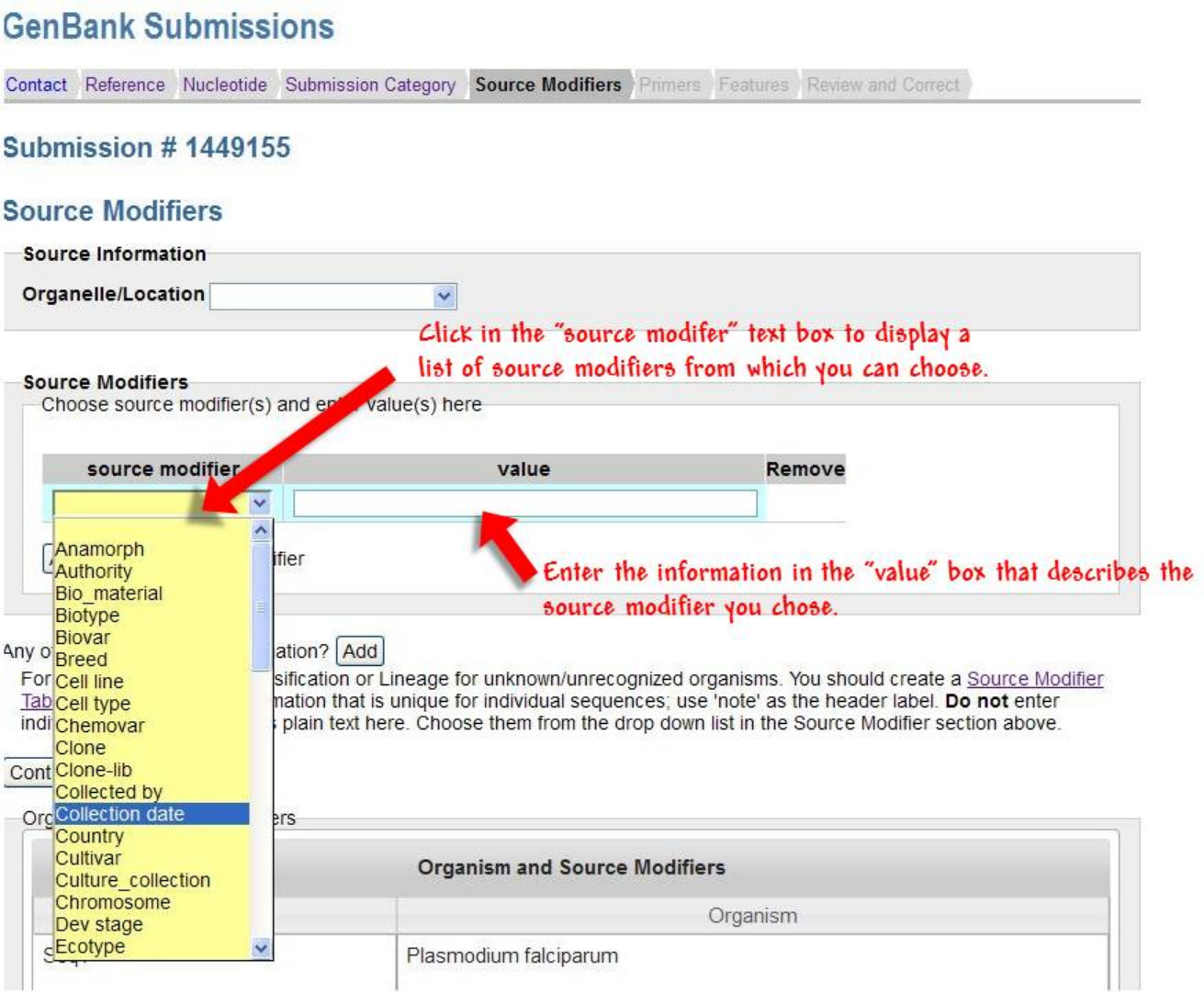

Figure 13: The "Source Modifiers" section of the "Source Modifiers" page when a single sequence is submitted. This view shows the drop-down list that appears when you click in the "source modifier" text box. Figure text gives a description of selecting the source modifier and entering information in the "value' text box.

and "value" text boxes will appear (Figure 16).

Use these boxes to enter the source in formation for your sequences. Clicking on the "source modifier" text box will release a drop-down menu of source modifiers you can select from (Figure 17). You can then enter the information that describes each source modifier in the "value" text box.

For example, if you had selected "collected by" as the source modifier, you would put the name of the person who collected the sample from which your sequence came in the "value" text box.

- For each additional source modifier you have, click the "Add" button to generate additional "source modifier" and "value" boxes you need to enter the information you have
- All of the source modifiers you enter will be applied to all of the sequences you submit.

Once you have entered your source modifiers, click the "Continue" button. This will update the static table at the bottom of the page with your source modifier selections. Review the table to make sure your source modifiers are correct, and then click the "Continue" button again to move to the next page of the form.

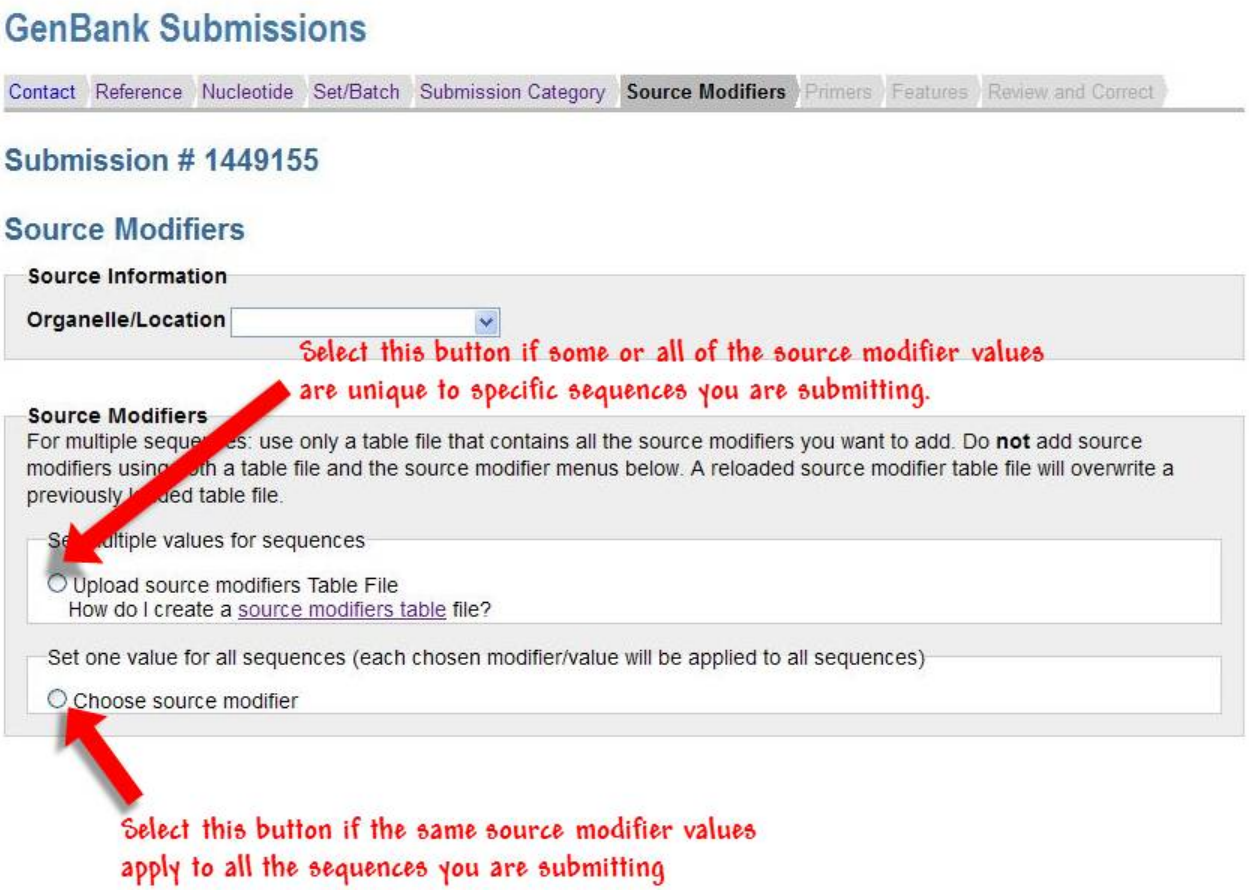

Figure 14: The "Source Modifiers" section of the "Source Modifiers" page when multiple sequences are submitted. Figure text gives information about which source modifier loading option to use.

Common Mistakes Made While Filling Out the "Source Modifiers" Page

- **Mistake: Using the wrong format for the collection date**

  Fix: The "collection date" source modifier needs to be in this format: dd-Mon-yyyy. An example of the correct format is 15-Dec-2011. If you enter the collection date using any other format, you will not be allowed to continue with your submission until you submit the collection date in the correct format.
- **Mistake: Using invalid source modifiers as column heads for your source modifier table**

  Fix: Use only valid source modifiers as column heads for your table. You can find a complete list of valid source modifiers on the BankIt source modifier help page.
- **Mistake: Uploading a Source Modifier table that is not formatted in plain text.**

  Fix: Since BankIt can read only plain text (.txt) files, if you upload a source modifier table in a format other than plain text, it will not load properly and BankIt will display an error message. You will not be

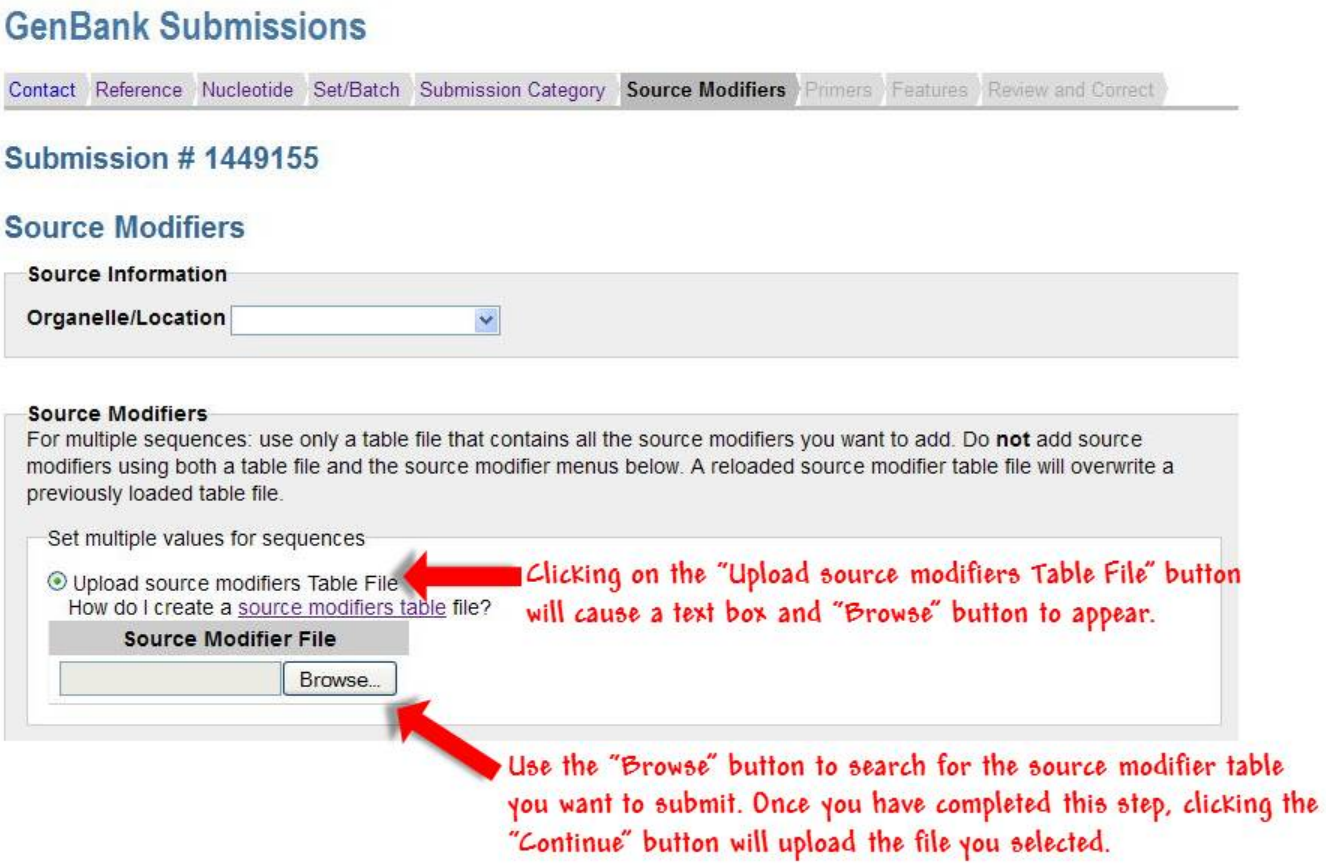

Figure 15: The "Source Modifiers" section of the "Source Modifiers" page when multiple sequences are submitted. This view shows the "Browse" button that appears when you click the "Upload source modifiers Table File" button in the "Set multiple values for sequences" sub-section. Figure text gives a description of selecting and uploading the source modifier table file.

allowed to proceed with your submission until you upload a correctly formatted source modifier table.

See the step-by-step instructions for making a tab-delimited source modifier table for information on making a table file in plain text (.txt) format, or use the BankIt Help documentation.

- **Mistake: Uploading a Flatfile list of Source Modifiers**

  Fix: Flatfile lists cannot be read by BankIt. If you upload a flatfile list of source modifiers, it will not load properly and BankIt will display an error message. You will not be allowed to proceed with your submission until you upload a tab delimited source modifier table in plain text format.

  See the step-by-step instructions for making a table file in plain text (.txt) format, or use the BankIt Help documentation.

  See the step-by-step instructions for making a tab-delimited source modifier table or use the BankIt Help documentation.

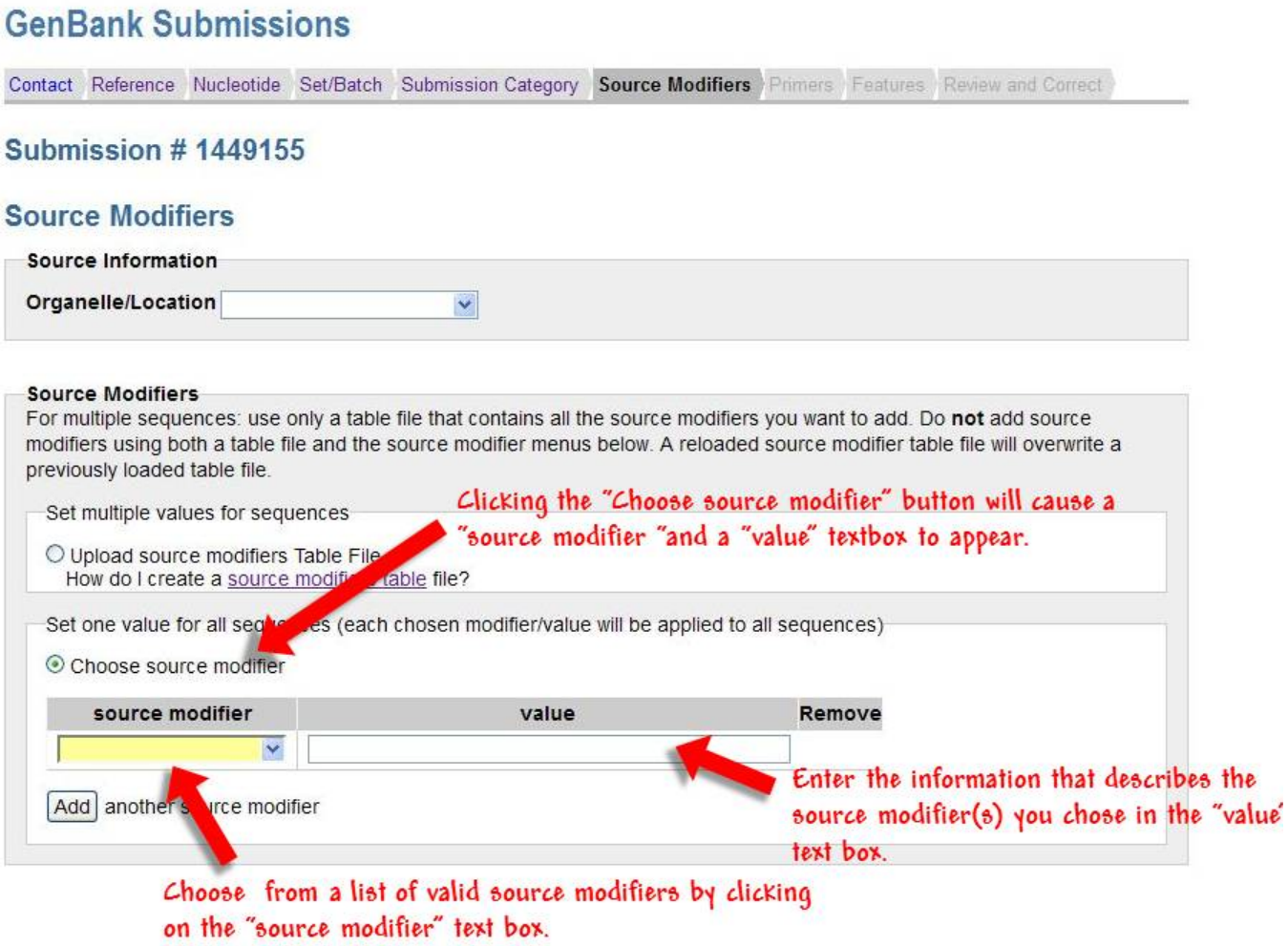

Figure 16: The "Source Modifiers" section of the "Source Modifiers" page when multiple sequences are submitted. This view shows the source modifier and value text boxes that appear when you click the "Choose source modifier" button in the "Set one value for all sequences" sub-section. Figure text gives a description of selecting the source modifier and entering information in the "value' text box.

Below are examples of flatfile format source modifiers lists that **cannot** be uploaded as source modifier files:

source 1..1510
/organism="Gallus gallus"
/mol_type="mRNA"
/breed="Roaster"
/chromosome="Z"
/map="Zp21"

source 1..549
/organism="Passiflora sprucei"
/organelle="plastid:chloroplast"
/mol_type="genomic DNA"

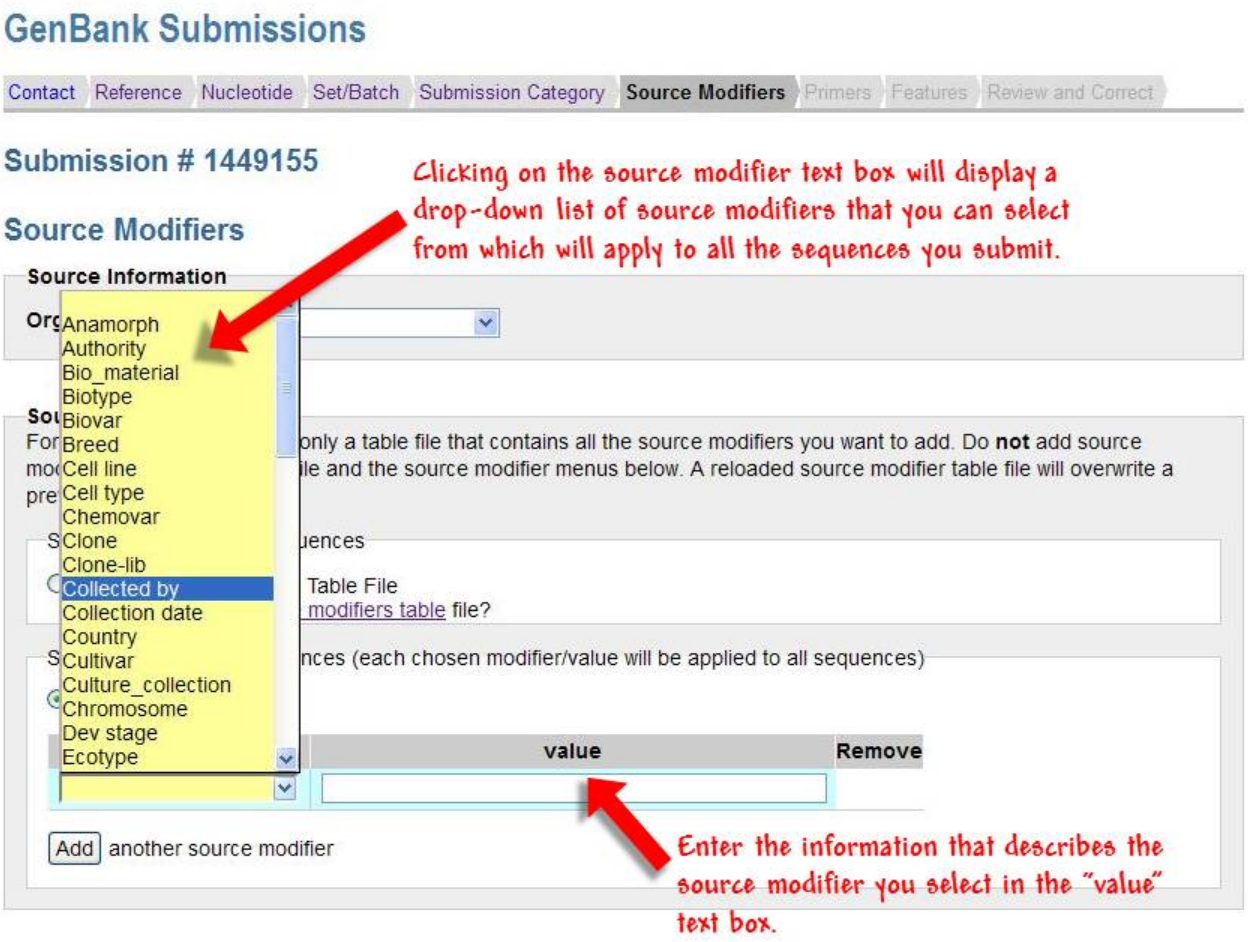

Figure 17: The "Source Modifiers" section of the "Source Modifiers" page when multiple sequences are submitted. This view shows the drop-down list that appears when you click in the "source modifier" text box in the "Set one value for all sequences" sub-section. Figure text gives a description of selecting the source modifier and entering information in the "value' text box.

/country="USA: Okanga National Forest"
/collected_by="John Jones"

- **Mistake: Pasting a flatfile list of source modifiers as an additional or corrected source modifier file on the "Review and Correct" page or as the only source information under 'Any other source related information?' on the Source Modifier page**

  Fix: These two text boxes allow a submitter to enter additional or corrected source modifier files or other plain text description for sequence source organisms after clicking a check box. Once you click on the check box, a text box will appear.

  A flatfile format list of source modifiers will be accepted in these text boxes because they are not restricted to the multi-column, tab-delimited format required for source modifiers; they will accept any text you put in them. However, your submission will not be accepted because we cannot upload flatfile format source modifier lists, and you will be requested either to:

- Resubmit your sequences in a new submission with the source modifier correctly input

  OR
- Send us your source modifier information in a tab-delimited plain text file.

The "Primers" Page

Michael Fetchko and Adrienne Kitts

Purpose

The BankIt submission tool "Primers" page is where you will provide the PCR primers you used to *amplify* the nucleic acid that you sequenced. The information we ask for in this page includes:

- The sequence of the forward and reverse PCR primers for each primer reaction set
- The names of the forward and reverse PCR primers for each primer reaction set

Do NOT submit sequencing primers in this page. This page is for entering the PCR primers used to amplify the nucleic acid that you sequenced.

Note: PCR primers are optional and are not required for a submission.

Single Sequence Submissions

If you submitted a single sequence in the "Nucleotide" page of the form, when you get to the "Primers" page, you will see primer sequence examples followed by text/name boxes for your primers (Figure 18):

Starting with the text box on the far left:

1. Place the **sequence** of your forward (fwd) primer in the **first** box.
2. Place the **name** of the forward primer in the **second** box.
3. Place the **sequence** of your reverse (rev) primer in the **third** box.
4. Place the **name** of the reverse primer in the **fourth** box.

Note: The primers you enter in these text boxes are PCR primers for nucleic acid amplification, and NOT sequencing primers.

Additional Primer Pairs for the Same Reaction Set

If you have additional pair(s) of forward and reverse primers for your PCR reaction mixture, click the "Add" button to get text boxes for another pair of forward and reverse primers (Figure 19). You can click the "Add" button as many times as you have primer pairs for this reaction mixture.

Additional PCR Reaction Sets for Your Sequence

- If you have another PCR reaction for the sequence you are submitting, click the "Add Another Primer Reaction Set" to get a set of boxes where you can add the forward and reverse primers for the additional reaction set (Figure 20).
- You can click the "Add" button to get text boxes for another primer set for this new reaction. You can click the "Add" button as many times as you have primer pairs for this reaction mixture.
- Once you have finished entering the primer pairs for this new reaction, if you have any additional reaction mixtures for this sequence, click the "Add Another Primer Reaction Set" button as many times as you have PCR reaction sets for this sequence that you need to add.

Check Your Primer Sequences before Continuing

Once you have finished entering the PCR primers for the sequence you are submitting, click the "Continue' button" to refresh the static table at the bottom of the "Primers" page with the sequences you have entered. Once

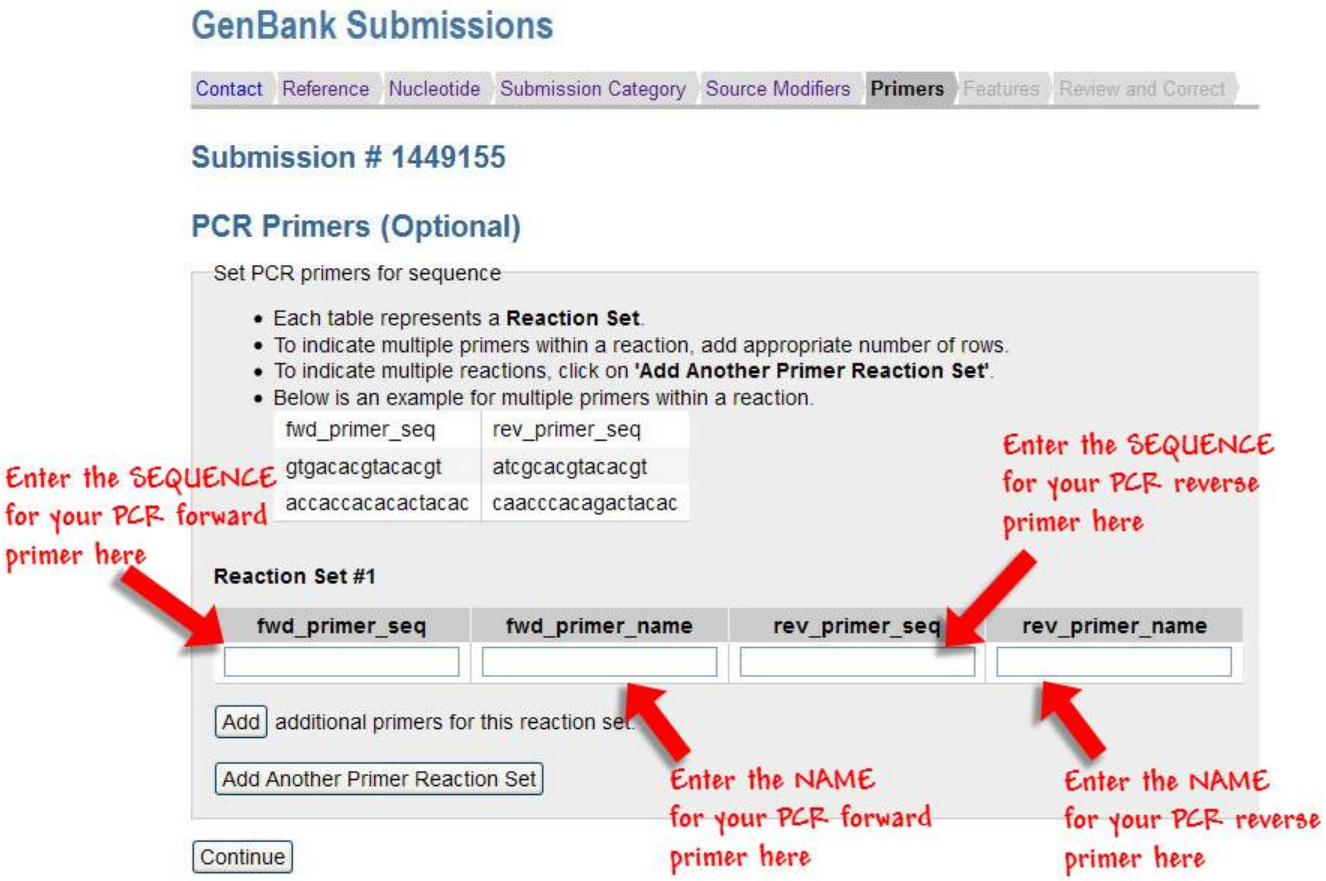

Figure 18: The "Primers" page when a single sequence is submitted. Figure text gives instructions for entering primer sequences and names.

the table has refreshed, check your primer sequences to be sure they are correct. Once you have verified that your primer sequences are correct, click the "Continue" button again to go to the next page of the form.

Multiple Sequence Submissions

If you submitted multiple sequences in the "Nucleotide" page of the form, when you get to the "Primers" page, you will be given a choice of uploading a Primers Table File or using the BankIt form to enter your primers. (Figure 21)

If the primer sets are different for each of the sequences:

1. Select the "Upload Primers Table File" button. When you do, a text box and "Browse" button will appear.
2. Use the "Browse" button to select the Primers Table File from your local computer.
3. Click the "Continue" button to load your Primers Table File.
4. The primers you submitted in your Primers Table File will be displayed in the "Primers" table at the bottom of the page.

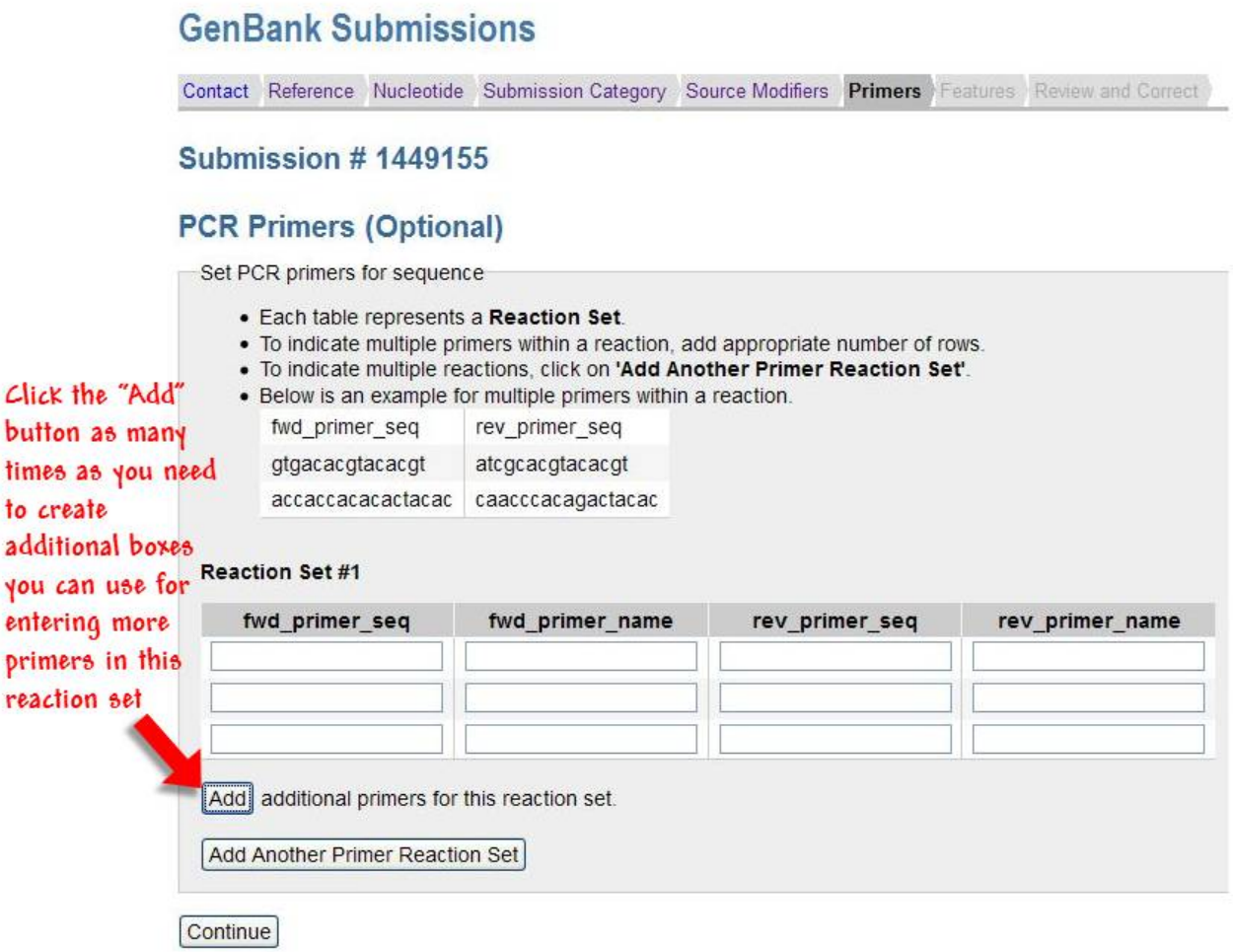

Figure 19: The "Primers" page when a single sequence is submitted and the "Add" button is clicked twice. Figure text gives instructions for using the "Add" button to enter more primers in a reaction set.

5. Check the Primers table at the bottom of the page to verify that the primer sequences displayed are correct.
6. Click the "Continue" button to go to the next page of the form.

Instructions for creating a Primers Table File are available from BankIt Help.

If the Primer Sets are the Same for all the Sequences:

Select the "Input PCR Primers" button. When you do, primer sequence examples followed by text/name boxes for your primers will appear (Figure 22):

Starting with the text box on the far left:

1. Place the **sequence** of your forward (fwd) primer in the **first** box.
2. Place the **name** of the forward primer in the **second** box.
3. Place the **sequence** of your reverse (rev) primer in the **third** box.
4. Place the **name** of the reverse primer in the **fourth** box.

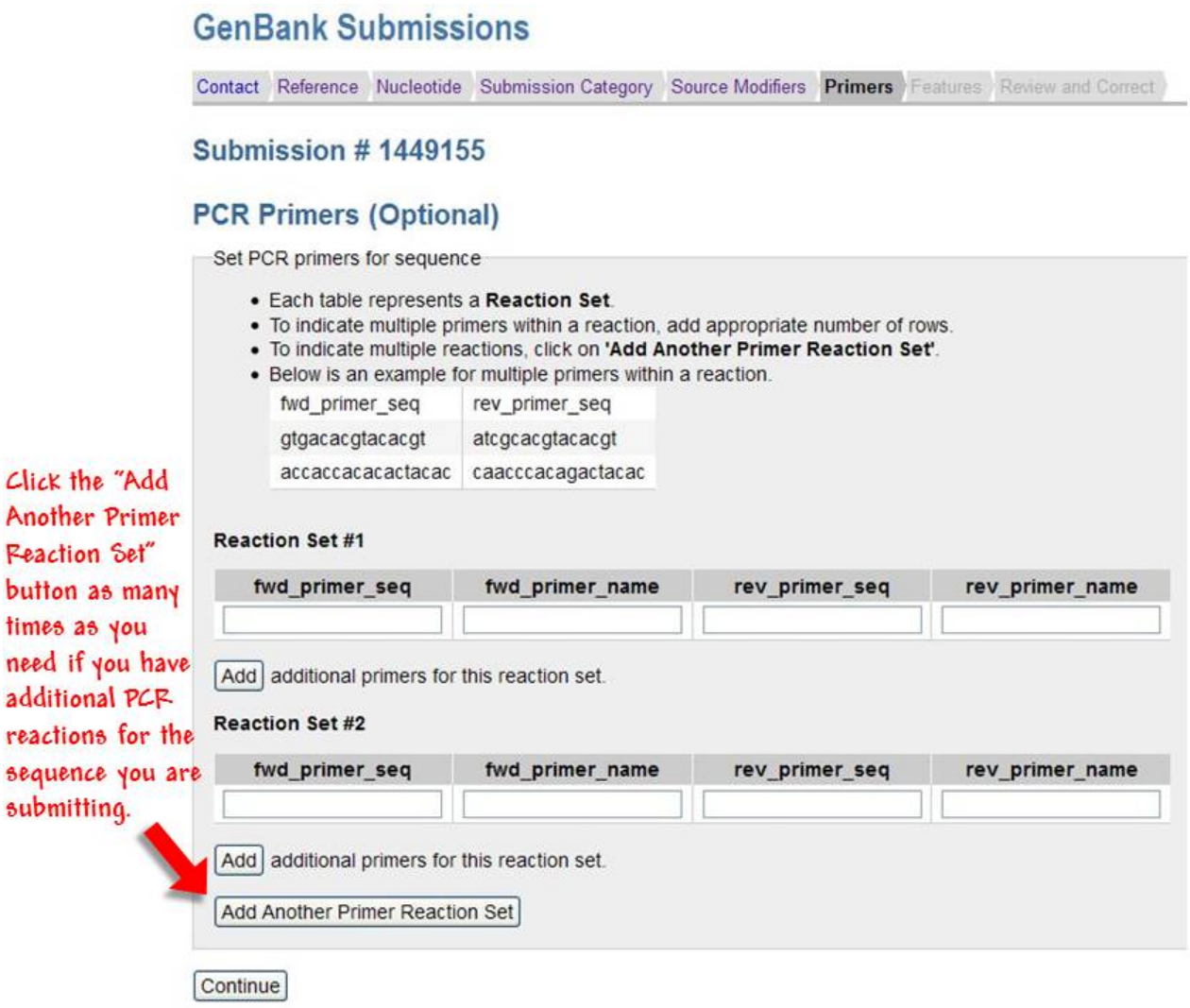

Figure 20: The "Primers" page when a single sequence is submitted and the "Add Another Primer Reaction Set" button is clicked once. Figure text gives instructions for using the "Add Another Primer Reaction Set" button to enter more PCR reaction sets for a sequence.

If you have additional primer pairs for the same reaction set, see the instructions in the single sequence submission section for entering additional primer pairs in the form.

If you have additional PCR reaction sets for your sequences see the instructions in the single sequence submission section for entering additional PCR reaction sets in the form.

Check Your Primer Sequences before Continuing

Once you have finished entering the PCR primers for the sequence you are submitting, click the "Continue" button to refresh the static table at the bottom of the "Primers" page with the sequences you have entered. Once the table has refreshed, check your primer sequences to be sure they are correct. Once you have verified that your primer sequences are correct, click the "Continue" button again to go to the next page of the form.

Common Mistakes Made While Filling Out the "Primers" Page

- **Mistake: Entering the primer sequences and names in the wrong text boxes.**

  Fix: Be sure to read the text boxes carefully before entering your data. Starting on the left, you enter the

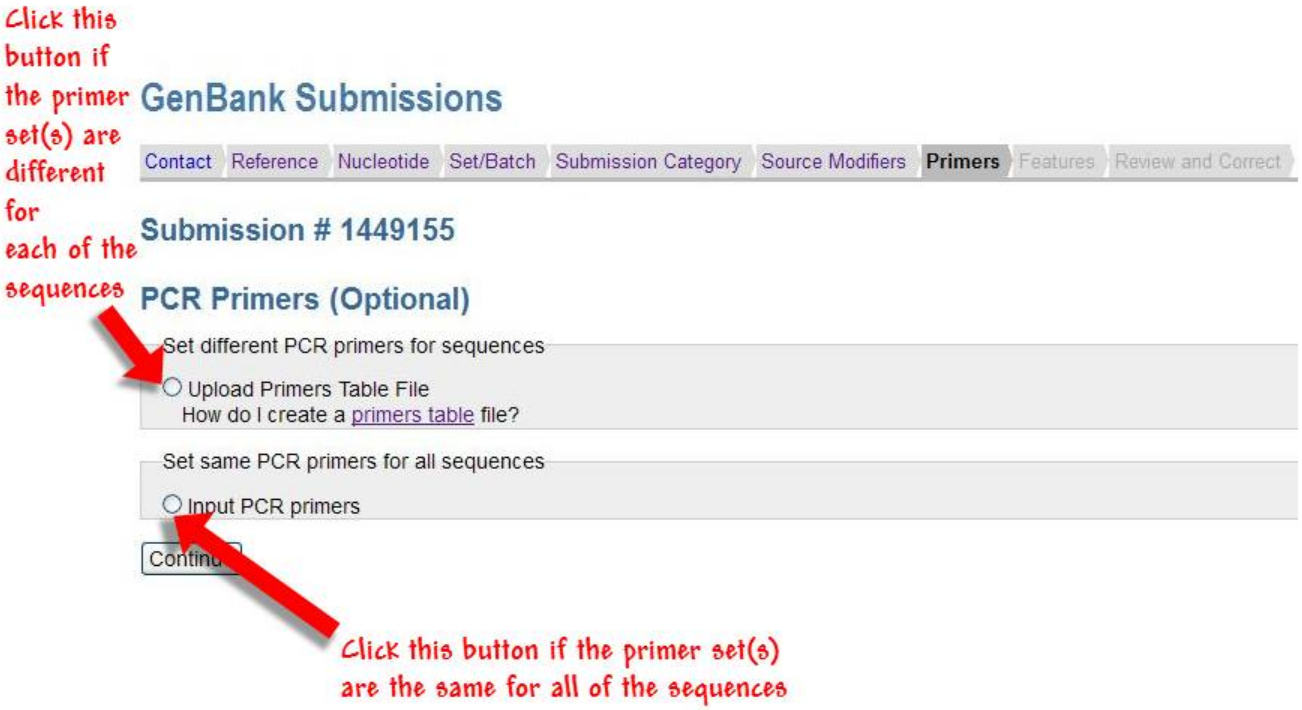

Figure 21: The "Primers" page when multiple sequences are submitted. Figure text gives hints for selecting a primer submission method.

forward primer sequence in the first box and the name for this primer in the second box. Then you enter the reverse primer sequence in the third box and the name for this primer in the fourth box.

- **Mistake: Entering sequencing primers instead of PCR primers.**

  Fix: This page is for entering PCR primers used to amplify the nucleic acid that you sequenced. Do not enter sequencing primers on this page.

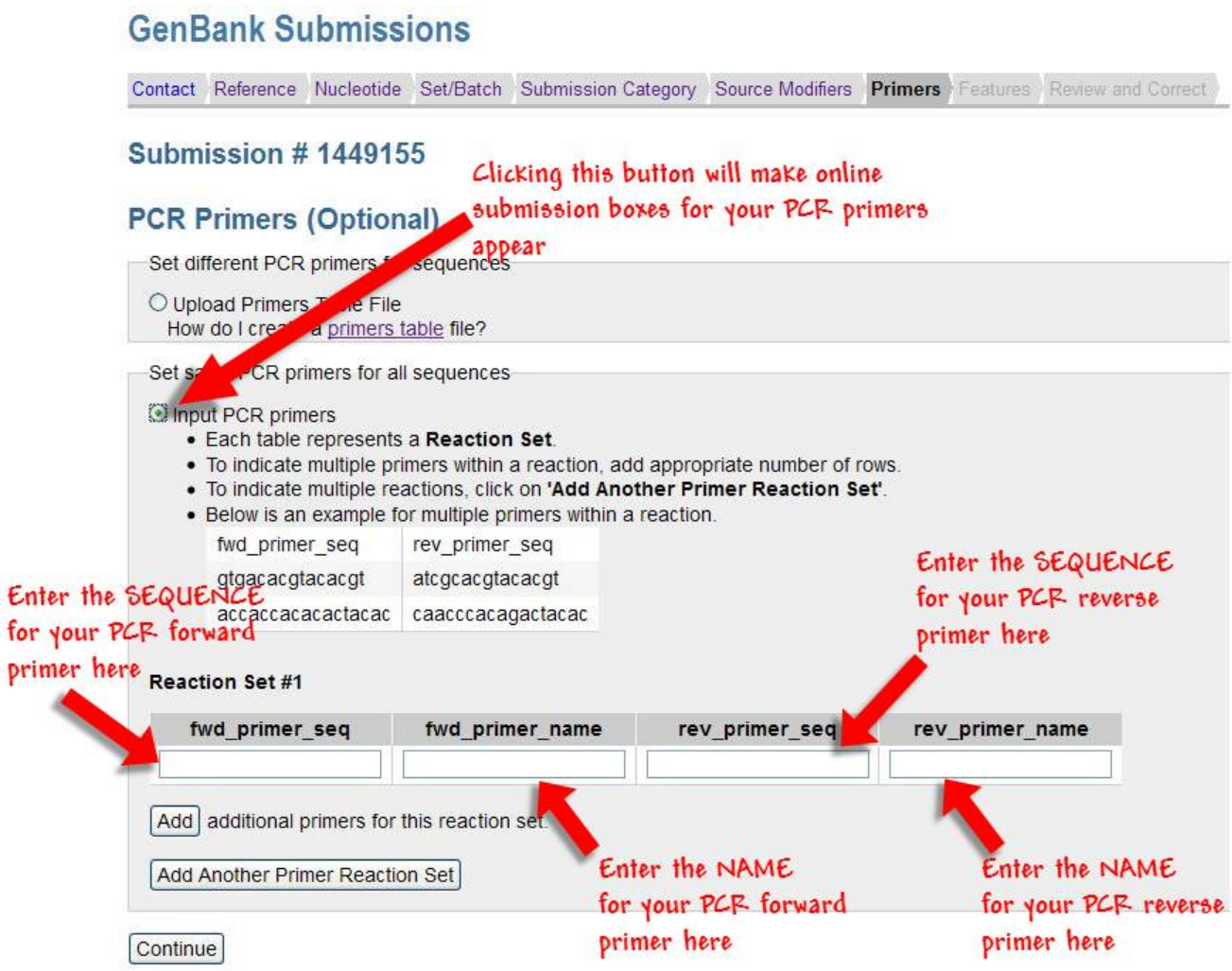

Figure 22: The "Primers" page when multiple sequences are submitted and the "Input PCR primers" button is selected. Figure text gives instructions for entering primer sequences and names.

The "Features" Page

Michael Fetchko and Adrienne Kitts

Purpose

The BankIt submission tool "Features" page is where you will select features for your sequence and provide detailed information about the features you selected. The detailed information we ask for in this section is dependent on the feature you select, but includes:

- The strand on which the feature appears
- Whether the feature is partial or complete. (If it is partial, if the feature is incomplete at the 5' end or at the 3' end)
- Whether the nucleotide interval where the feature occurs spans the entire sequence or has a specific span. If it has a specific span you must provide the nucleotide numbers of that span.
- Qualifiers for the feature

Adding Features to your Sequence: Feature Table File vs. Online BankIt Forms

Uploading a feature table file is an efficient method of adding features to sequence(s) if:

- you are adding many different features to a single sequence
- you are adding many different features to a number of different sequences

Complete the BankIt feature form(s) if:

- you are adding a one or a few features to a single sequence submission
- you are adding the same features to all the sequences in a multi-sequence submission
- you are adding the same features to specific sequences in a multi-sequence submission

Adding Features by Uploading a Feature Table

What is a five column feature table and how do I make one?

- A tab-delimited feature table uses a single "Tab" keystroke to delimit (mark the boundary) between one column and the next in a table that contains your feature information.
- BankIt Help documentation contains information about how to format a feature table and provides examples.
- You can access step-by-step instructions for creating a feature table in the GenBank Submission Resources Quick Start.
- A list of valid Features and Qualifiers you can use in your table are available in the BankIt Help documentation using links found within BankIt's "Feature" pages.
- A Feature Table file must be saved in plain text format

  Note: If you upload a table that includes invalid features or qualifiers, BankIt will tell you to correct your table with valid features/qualifiers and reload it.

See Figure 23 for a sample feature table marked up to show where to place your features and feature modifiers in a feature table.

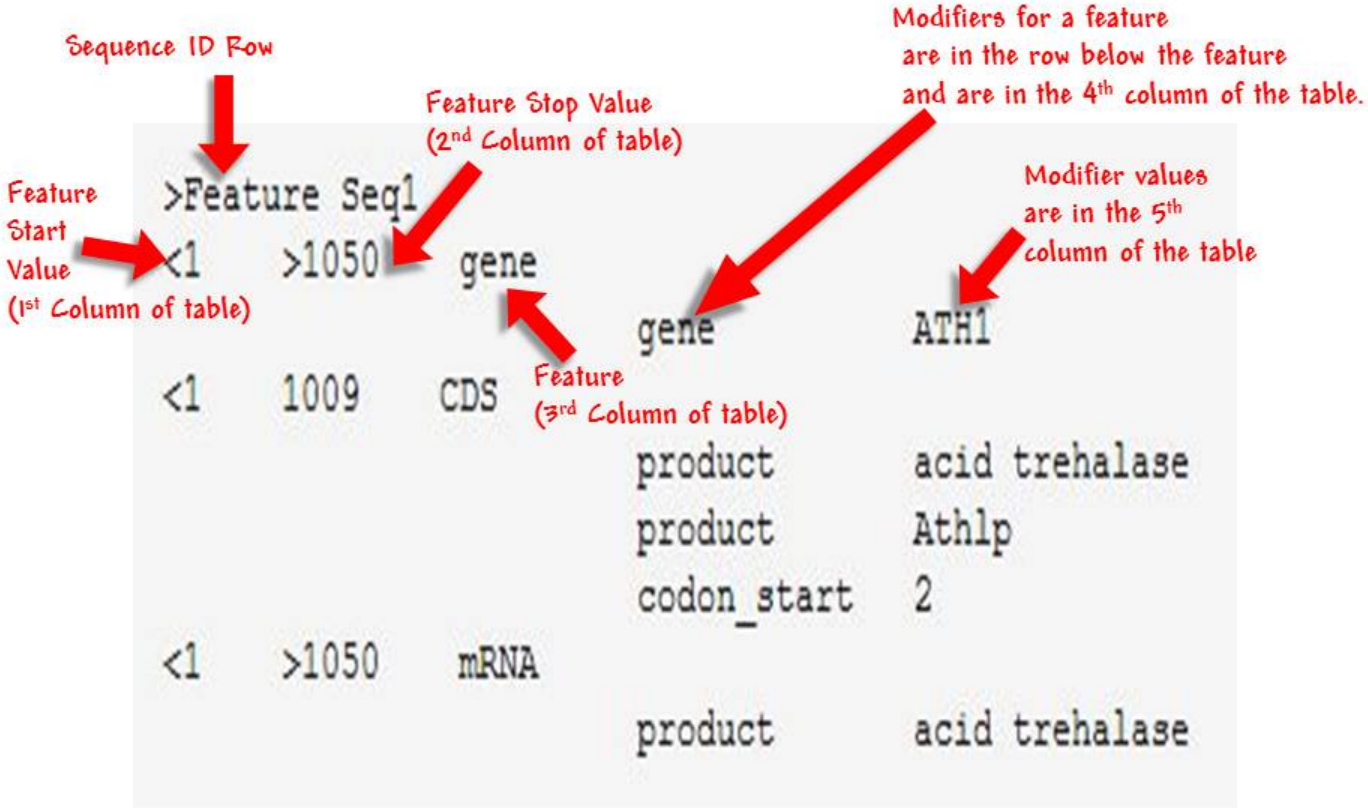

Figure 23: A sample feature table. Figure text shows the correct location for features and feature modifiers.

The Importance of Formatting your Feature Table Correctly

The 5 column feature table format is specific so that BankIt can read the information in the table and put it in the right place within your submission. It is important that you follow the feature table format instructions and examples we provide. If your table differs from the format provided in the instructions and examples, BankIt will not be able to read it, and you will be requested to correct your table and upload it again.

Even if your table strays from the accepted feature table format in a very small way — like using a space between the columns instead of a tab, BankIt will not be able to read your table.

If BankIt does not accept your feature table, recheck your table carefully to see where it differs from the accepted format.

The Feature Table must be saved in a plain text format.

Uploading your Feature Table

Once you select the "Add features by uploading five column feature table file" option, a text box and a "Browse" button will appear on the "Features Overview" page. Click the "Browse" button and select the feature table file that you created and saved on your computer. Then click the "Upload File" button to upload the file.

Once you have uploaded the file, BankIt will read it and generate a list of the features that are in your table followed by a display of how those features will appear in your sequence record at the bottom of the page.

If there are errors in the feature table file, BankIt will display error and/or warning messages that describe the problem and direct you how to fix it.

If you review the table, and find that you need to make some changes, use the "Edit" and "Remove" buttons in the feature list to update the features you added. Once you are satisfied with the features you added, Click the "Continue" button to move on to the next page of the form.

Adding Features using the Online BankIt Forms

Once you select the "Add features by completing input forms" option, a list of 4 feature categories will appear on the "Features Overview" page:

- CDS
- RNA
- Repeat
- Other

When you select one of the categories, you may be directed to select a feature using a drop-down list or a button activated choice. Select one of the features from the drop-down menu or a button activated choice if they appear, then click the "Add" button to go to the "Features Detail" page for the feature. Once you are on the "Feature Details" page, you will give us detailed information about the feature you chose.

The Coding Region Feature Category

If the sequence you are submitting encodes a protein, select this feature category. When you do, options for how you will add the Coding Region (also called "coding sequence" or "CDS") feature to your sequence record will appear. Once you select the method by which you will add the CDS feature, click the "Add" button and use the online "Features Detail" page that appears to give us detailed information about the CDS feature you are adding to your submission. For more instructions on how to provide detailed information and qualifiers for the CDS feature, see Box 4.

Box 4: The "Coding Region" Category: Selecting a Feature and Providing Information for it.

1. **Select the "Coding Region" category**
2. **Select an option on how you will add the Coding Region**.
3. **Click the "Add" button**. A "Features Detail" page will appear for you to provide intervals or to provide protein sequence data.
4. **Answer the remaining questions in the "Feature Detail" page** with the general information requested for the feature.
5. **Click to activate the "Qualifier" drop down menu and select the qualifiers** for the "Coding region" feature you chose.
6. **Provide values for the qualifiers you chose**.

Box 4 continues on next page...

Box 4 continued from previous page.

7. **Click the "Add" button to add more qualifiers for your feature or click the "Accept" button** to record the feature details you entered into your submission.
8. **Clicking the "Accept" button will** make a list of the features you added appear followed by a display of how those features will appear in your sequence record.
9. **Review the feature display**.
10. **Go to the "Added Features for editing/removal" list** above the Features display to:
    a. Remove that feature from your record (click the "X" button located to the right of a feature to remove it).

    OR

    b. **Change the information you entered for this particular feature** (click the "Edit" button to go back to go back to the "Feature Detail" page where you can change the information you provided for the feature)
11. **Click the "Continue" button** to continue to the next page of the BankIt form once you are satisfied with the features you see in the "Features" display at the bottom of the page.

Adding the CDS feature by providing intervals

If you choose this option to add the CDS feature to your record, in addition to providing qualifiers and general information about the coding sequence and the protein it contains, you will need to provide the coding region spans on the sequence (e.g. multiple intervals if your sequence contains introns, or the entire sequence if it does not). If the coding region spans specific part(s) of the sequence, you will need to give us the nucleotide numbers within the sequence for each span of the coding region.

Adding the CDS feature by providing protein sequence data

If you choose this option to add the CDS feature to your record, you must provide the sequence of the protein encoded by the coding region by either uploading a protein FASTA file , or typing/pasting the sequence in the space provided in the "Features Detail" page.

Coding Region Feature Qualifiers

Once you get to the "Features Detail" page for the CDS feature, in addition to the other information requested on the page, you will also be asked to select feature qualifiers for the coding sequence. A list of valid Qualifiers you can use are available in the BankIt Help documentation using links found within BankIt's "Feature" pages.

A product (protein) name (or description) is required for each CDS. If a gene is added, you must also provide a gene name (or description).

The RNA Feature Category

The "RNA" feature category allows you to select an RNA feature for your sequence. Once you select the RNA feature category option and click the "Add" button, a drop-down menu will appear. Click in the text box of the drop-down menu to release a list of RNA features. Once you select the one of the "RNA" features from the drop-down menu and click the "Add" button, you will use the "Features Detail" page to select qualifiers for and give us more detailed information about the feature you chose. For more instructions on how to select "RNA" features and provide information and qualifiers for the "RNA" feature(s) you select, see Box 5.

Box 5: The "RNA" Category: Selecting a Feature and Providing Information for it.

1. **Select the "RNA" features category to make a dropdown menu appear**.
2. **Click on the drop down menu to release a list of RNA features**.
3. **Select a feature and click the "Add" button**. A "Features Detail" page will appear where you can provide the specific information about the feature you selected.
4. **Answer the questions in the "Feature Detail" page** with information about the feature you selected.
5. **Click to activate the "Qualifier" drop-down menu and select the qualifiers** for the "RNA" feature you chose.
6. **Provide values for the qualifiers you chose**.
7. **Click the "Add" button to add more qualifiers** for your feature **or click the "Accept" button** to record the feature details you entered into your submission.
8. **Clicking the "Accept" button will** make a list of the features you added appear followed by a display of how those features will appear in your sequence record.
9. **Review the feature display**.
10. **Go to the "Added Features for editing/removal" list above the features display to**:
    a. **Remove that feature from your record** (click the "X" button located to the right of a feature to remove it).

    OR

    b. **Change the information you entered for this particular feature** (click the "Edit" button to go back to go back to the "Feature Detail" page where you can change the information you provided for the feature)
11. Click the "Continue" button to continue to the next page of the BankIt form once you are satisfied with the features you see in the "Features" display at the bottom of the page.

RNA Feature Types and Definitions

- **premessage RNA**
  is an RNA molecule that was not processed after it was made and therefore contains intervening sequences (introns) in addition to the 5' untranslated region (5' UTR), the coding sequences (CDS, exon), and the 3' untranslated region (3' UTR) normally found in mature (processed) mRNA.
- **messenger RNA (mRNA)**
  is RNA that encodes a protein. An mRNA for a protein product includes the 5' untranslated region (5'UTR), the coding sequence (CDS, exon) and the 3' untranslated region (3'UTR).
- **transfer RNA (tRNA)**
  is a small RNA molecule (75-85 bases long) that facilitates the translation of a nucleic acid sequence into an amino acid sequence.
- **ribosomal RNA (rRNA)**
  is the RNA component of the ribonucleoprotein particle (ribosome) which assembles amino acids into proteins. (e.g. 16S rRNA, 28S rRNA, large subunit rRNA).
- **non-coding RNA (ncRNA)**
  is RNA that does not encode a protein. This feature should not be used for ribosomal RNA (rRNA) and transfer RNA (tRNA) as they have their own feature keys (examples of non-coding RNA include scRNA, snRNA, miRNA, and siRNA).
- **transfer messenger RNA (tmRNA)**
  acts as a tRNA first, and then as an mRNA that encodes a peptide tag. The ribosome translates the mRNA region of the tmRNA and then attaches the encoded peptide tag to the C-terminus of the unfinished protein. The attached tag targets the protein for destruction or proteolysis.

- **miscellaneous RNA (misc_RNA)**
  is any transcript or RNA product that cannot be defined by other RNA feature types listed

RNA Feature Qualifiers

Once you get to the "Features Detail" page for the RNA feature you select, you will be asked to select feature qualifiers for the RNA feature you chose. A list of valid Qualifiers you can use are available in the BankIt Help documentation using links found within BankIt's "Feature" pages.

Note: You must provide a product name or description for all RNA features.

The Repeat Region Feature Category

The Repeat region category allows you to choose a repetitive element feature for your sequence. Once you select the Repeat region category button and then click the "Add" button, you will use the online "Features Detail" page that appears to select qualifiers for and give us more detailed information about the feature you chose. For more instructions on how to select a Repeat region feature and provide information and qualifiers for it, see Box 6.

Box 6: The "Repeat region" Category: Selecting a Feature and Providing Information for it.

1. **Select the "Repeat Region" category**
2. **Click the "Add" button**. A "Features Detail" page will appear.
3. **Choose the specific type of repetitive element** (e.g. mobile element, satellite or repeat sequence) you are submitting.
4. **A "Type" dropdown menu will appear** from which you must choose the specific type of mobile element, satellite or repeat sequence you are submitting.
5. **Provide the name you gave to the specific repeat you are submitting**.
6. **Answer the questions in the "Feature Detail" page** with general information about the feature you selected.
7. **Click to activate the "Qualifier" drop down menu** and select the qualifiers for the "Repeat region" feature you chose.
8. **Provide values for the qualifiers you chose**.
9. **Click the "Add" button to add more qualifiers for your feature or you click the "Accept" button** to record the feature details you entered into your submission.
10. **Clicking the "Accept" button** will make a list of the features you added appear followed by a display of how those features will appear in your sequence record.
11. **Review the feature display**.
12. **Go to the "Added Features for editing/removal" list** above the Features display to:
    a. **Remove that feature from your record** (click the "X" button located to the right of a feature to remove it).

    OR

    b. **Change the information you entered for this particular feature** (click the "Edit" button to go back to go back to the "Feature Detail" page where you can change the information you provided for the feature)
13. **Click the "Continue" button** to continue to the next page of the BankIt form once you are satisfied with the features you see in the "Features" display at the bottom of the page.

Repeat Region Feature Definitions

- **Repeat Sequence**
  is a specific nucleotide sequence (unit) that recurs multiple times in a genome. A repeat unit can be arranged in any of the following ways:
  - **Tandem repeat**
    is a repeating nucleotide sequence that exists end-to-end in the same orientation with another copy of that nucleotide sequence.
  - **Inverted repeat**
    is a repeating nucleotide sequence that normally occurs as part of an end-to-end pair. The first member of the pair is the repeating nucleotide sequence oriented in the forward direction. The second member of the pair is the repeating nucleotide sequence oriented in the reverse direction.
  - **Flanking repeat**
    is a repeating nucleotide sequence that lies outside the sequence for which it is functionally important (e.g. transposon insertion target sites).
  - **Terminal repeat**
    is a repeating nucleotide sequence that occurs both:
    - At the ends of sequence for which it is functionally important

      AND
    - Within sequence for which it is functionally important
  - **Direct repeat**
    is a repeating nucleotide sequence that does not always lie end-to-end with another copy of that nucleotide sequence, but is in the same orientation with it.
  - **Dispersed repeat**
    is a repeating nucleotide sequence that is found scattered throughout the genome.
  - **Other repeat**
    is a repeating nucleotide sequence with important characteristics that are not described by the other repeat types listed above.
- **Satellite DNA**
  is made of many tandem repeats (identical or related) of a short, basic nucleotide sequence. It is frequently found in the centromere of a chromosome, but can also be found elsewhere. Because of its base composition, satellite DNA's density is such that it will form bands in a CsCl buoyant density gradient that are "satellite" (separate from but close to) the bands formed by genomic DNA.
- **Mobile Element**
  - *is a genetic entity**that is capable of movement**from one location to another in the genome.*

    OR
  - *is a genetic entity**that is derived from the movement**from one location to another in the genome.*

Repeat Region Feature Qualifier Definitions

Once you get to the "Features Detail" page for the Repeat region feature you select, you will be asked to select qualifiers for the repeat region feature you chose. A list of valid Qualifiers are available in the BankIt Help documentation using links found within BankIt's "Feature" pages.

The "Other" Feature Category

The "Other" feature category allows you to select features for your sequence that are not shown on the "Feature Overview" page. Once you select the "Other" feature option and click the "Add" button, a drop-down menu will

appear. Click on the drop-down menu to display a list of "Other" features. Once you select one of the "Other" features from the drop-down menu and click the "Add" button, you will use the online "Features Detail" page that appears to select qualifiers for and provide more detailed information about the feature you chose. For more instructions on how to select "Other" features and provide information and qualifiers for the "Other" feature(s) you select, see Box 7.

Box 7: The "Other" Category: Selecting a Feature and Providing Information for it.

1. **Select the "Other" features category to make a drop-down menu appear.**
2. **Click on the drop down menu** to release a list of features not found on the "Feature Overview" page.
3. **Select a feature and click the "Add" button.** A "Features Detail" page will appear where you can provide the specific information about the Feature Category you selected.
4. **Answer the questions in the "Feature Detail" page** with information about the feature you selected.
5. **Click to activate the "Qualifier" drop-down menu and select the qualifiers** for the "other" feature you chose.
6. **Provide values for the qualifiers you chose.**
7. **Click the "Add" button to add more qualifiers for your feature or click the "Accept" button** to record the feature details you entered into your submission.
8. **Clicking the "Accept"** button will make a list of the features you added appear followed by a display of how those features will appear in your sequence record.
9. **Review the feature display**.
10. **Go to the "Added Features for editing/removal" list above the Features display to**:
    a. **Remove that feature from your record** (click the "X" button located to the right of a feature to remove it).

    OR

    b. **Change the information you entered for this particular feature** (click the "Edit" button to go back to go back to the "Feature Detail" page where you can change the information you provided for the feature)
11. **Click the "Continue" button to continue to the next page of the BankIt form** once you are satisfied with the features you see in the "Features" display at the bottom of the page.

"Other" Feature Definitions

You can see definitions for the features in the "Other" feature category in the BankIt Help documentation links in the "Features" page.

"Other" Feature Qualifier Definitions

Once you get to the "Features Detail" page, you will be asked to select feature qualifiers for the "Other" feature you chose. A list of valid Qualifiers are available in the BankIt Help documentation using links found within BankIt's "Feature" pages.

Common Mistakes Made While Filling Out the "Features" Page

Mistake: Selecting the "Coding Region" feature category and providing only gene information in the "Features Detail" page.

Fix: If you select the "Coding Region" feature category and provide only your gene information and not the actual coding region, BankIt will display an error message that additional information for the CDS is required.

When you select the "Coding Region" category, be sure to provide either the nucleotide interval spans for the CDS or the actual protein sequence data in addition to the product (protein) name and the gene information.

Mistake: Choosing only "Gene" or "Exon" for a sequence that encodes a protein.

Fix: You should select the coding region (CDS) category to add a correct CDS feature. If you do not, your submission may be returned to you.

Mistake: Creating a feature table using features and qualifiers that are not valid.

Fix: Your table must include only those features and qualifiers found in the valid "Feature" and "Qualifier" lists that are linked to from the Bankit Help Documentation.

Mistake: Creating a feature table that does not follow the feature table format provided in the instructions and examples.

Fix: The 5 column feature table format is very specific – BankIt can only read a feature table that follows this format exactly. If your table differs from the format provided in the instructions and examples, BankIt will not be able to read it, and you will be requested to correct your table and upload it again.

Even if your table differs from the accepted feature table format in a small way, BankIt will not be able to read your table.

If BankIt does not accept your feature table, recheck your table carefully to see where differs from the accepted format.

The "Review and Correct" Page

Michael Fetchko and Adrienne Kitts

Purpose

The BankIt submission tool "Review and Correct" page is where you will review the preliminary flatfiles BankIt generates from the data you entered into the BankIt tool. If you find that you made some mistakes in your data entry or you left out some information, you can use the page tabs to go back to any section in the form and correct the data you entered there. Once you have corrected your data, return to the "Review and Correct" page and verify that the changes you made are in the flatfiles BankIt generates from your corrected data.

Note: If you navigate to the Nucleotide page and alter the sequences on that page, you will be warned that you will lose any Source, Feature, or other descriptors that you have already input. If you proceed in making changes to the sequences on the Nucleotide page, you will have to re-enter the Source, Feature and other descriptive information for that sequence.

Correspondence

The "Review and Correct" page is also where you will verify the email address(es) of the person(s) we send our correspondence to regarding your submission. If there is a person we should contact about your submission in addition to the person listed on the 'Contact" page, you can enter their email address in the text box provided.

If You Have Been Asked to Resubmit Your Sequence(s)

If you have been asked to resubmit your sequence data because they were missing information, click the check box that follows the text: "If you have additional or corrected source modifier or feature files, or other plain text description for your sequence data submission, check here". When you click the box, a text box will appear in which you should type "resubmission of BankIt ####### "(without quotes), where ######## is the BankIt ID# of the previous submission you have been asked to resubmit.

Additional Text Descriptions of your Sequence

If you have additional information about your sequence that didn't fit into the answers you gave to the questions in the previous pages, click the check box that follows the text: "If you have additional or corrected source modifier or feature files, or other plain text description for your sequence data submission, check here".

When you click the box, a text box will appear which you can use to type or paste any additional descriptive text you have for your sequence.

Additional/Corrected Source Modifier/Feature Tables

If you made a correction or addition to a source modifier or feature table and haven't yet uploaded to BankIt, click the check box that follows the text: "If you have additional or corrected source modifier or feature files, or other plain text description for your sequence data submission, check here".

When you click this box, a text box and "Browse" button will appear for you to use to select the files you wish to upload. The new or corrected files will be submitted with your BankIt submission when you click the "Finish Submission" button.

Note: We prefer that you upload the correct files/tables at the time of submission on the correct page. If we cannot use a corrected file/table that you upload on the "Review and Correct" page, we may ask you to resubmit your data.

Reviewing your Submission

At the bottom of the "Review and Correct" page you will find the preliminary flatfiles that BankIt generated from the data you entered into the BankIt form. Review these files for mistakes you may have made or for data you may have forgotten to put in the submission form.

If you have submitted multiple sequences, be aware that only the first 5 flatfiles of your submitted sequences. You can be confident that BankIt captured all of the sequences you submitted; 5 sequences are a good sample of a multiple sequence submission to review to be sure you entered everything correctly.

If you submitted more than 5 sequences, and want to view them all, click the link located just above the "Finish" Submission" button in the "Review Records" section of the page that will allow you to download your submission as a ZIP file to your computer.

Submitting Your Sequence

Once you have reviewed the BankIt Flatfile for your submission and corrected and/or added any new data, click the "Finish Submission" button. When you do, you will see a "Submission Completed" message that will confirm your BankIt submission ID.

Submission ID vs. Accession Number

The BankIt submission ID is **not** an accession number. GenBank accession numbers will be assigned to your sequences and sent to you by email within two working days.

The time it takes to receive your accession number(s) may be longer if there is issues with your submission that need to be resolved. In that case, we will contact you, and once the issue is resolved, we will assign an accession number.

Emails Sent by BankIt Once you Submit

For each complete sequence submission, you will receive the following at the email address you provided in your submission:

1. An automatic reply confirming our receipt of your submissions. This reply is generated immediately after you make your submission (depending on your email system, it may take several hours to arrive at your email address).
2. GenBank accession numbers for all submitted sequences (they should arrive within two working days) unless there are problems with your submission that we must first resolve with you.
3. Your final GenBank record(s) for you to review before we release them to the Public database. These records have been processed by the GenBank Annotation staff and incorporate all the information you provided in your submission.

If has been two working days (Monday – Friday) since you completed the BankIt submissions process, and you haven't yet received a response from us following the automated confirmation of your submission, email a message to gb-admin@ncbi.nlm.nih.gov that asks us to check the status of your submission. Be sure to include in the message:

- The email address you used in your submission to GenBank
- The BankIt submission ID provided in our automated submission response
- The date you completed the BankIt submission process.

After Submission

Michael Fetchko and Adrienne Kitts

Submission Processing

Submissions are not automatically deposited into the GenBank database after being assigned their accession numbers.

Your sequences will first be examined and processed individually by the GenBank annotation staff members to determine if they contain errors or problems.

When your record is processed, we will contact you if we require additional information.

When your record is complete, a final copy of your GenBank record will be sent to you, and the record will be made publicly available on the date you requested.

Contacting GenBank about a Submission you have made

If you have any questions or corrections regarding your submission(s) contact gb-admin@ncbi.nlm.nih.gov . Be sure to include:

- The accession number for the record in question
- The email address used in your submission.

Accessing your Submission File Once you Submit

If you want to access your submission file once you have completed your submission, you can always download a ZIP file of your submission record from the “Submissions” page of your BankIt account. The “Submissions” page is the first page to appear once you login to use BankIt.

Note: You may not use BankIt to update a completed submission.

Updating your Submission

You may update or revise your submissions at any time by sending new or corrected information in an email to update@ncbi.nlm.nih.gov . You may also contact us at this address with any questions.

See the GenBank Update page for information on how to update an existing submission/record.

A User's Guide to the Sequin Wizards

Created: May 15, 2012; Updated: January 16, 2014.

The Design of this User's Guide

Created: May 15, 2012; Updated: January 16, 2014.

This User's Guide is designed to be a practical, plain language guide for the beginner— or for anyone who wants basic instructions for any of the steps you must perform when you submit using the Sequin Wizards.

Each major section in this User's Guide provides instructions for using each Sequin Wizard Submission Tool. In each section you will find a summary of purpose for each wizard and step-by-step instructions to guide you through the steps in the wizard.

What are the Sequin Wizards?

Created: May 15, 2012; Updated: January 16, 2014.

The Sequin Wizards will guide you in preparing your submission file for specific types of sequence submissions in the NCBI Sequin Submission Tool. The wizards are designed to capture all the required source information and will provide some assistance with annotation. Currently, there are wizards available for the following types of submissions:

- Viruses
- Uncultured Samples
- rRNA-ITS-IGS Sequences
- Intergenic Spacer Sequences
- Microsatellite Sequences
- D-Loop/Control Region Sequences

Prior to using the Wizards, make sure you have:

- Removed all contaminating vector sequence
- Quality trimmed your sequences
- Checked 16S ribosomal RNA submissions for anomalies using a chimera check tool. Anomalous sequences should be removed from your FASTA file.

How do you access the Wizards?

Created: May 15, 2012; Updated: January 16, 2014.

The Wizards are accessed in the NCBI Sequin Submission Tool. Please download the latest version of Sequin from NCBI to access the submission wizards described in this User Guide.

After you start a new submission and fill in the author forms, you will be prompted to select how you want to prepare your submission in the Preparing the Sequences dialog (Figure 1).

To begin, select the appropriate radio button under the section labeled Use a Submission Wizard. If you have selected the normal submission dialogs, you may be prompted to use a wizard if you have provided information suggesting it is a certain submission type.

If you have questions about using the Sequin Wizards, contact GenBank User Services at: info@ncbi.nlm.nih.gov

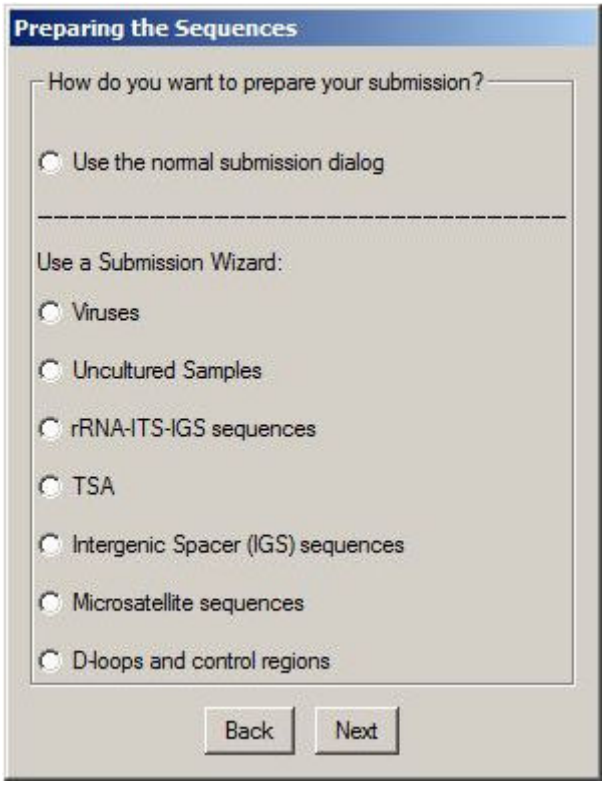

Figure 1. Preparing the Sequences dialog for selecting a Sequin Wizard.

Submission Wizard for Viruses

Created: May 15, 2012; Updated: January 16, 2014.

Purpose

The Virus Sequence Submission Wizard is for submitting virus and viroid sequence submissions only. This wizard will guide you in providing all of the necessary source information for different types of viruses and will provide assistance and direction with feature annotation. Examples of source information are provided.

Wizard Import Nucleotide Sequences

Requirements: The Virus Sequence Submission requirements are listed in the Sequences tab of the Wizard Import Nucleotide Sequences dialog box.

Sequence Format: You may import your sequences in FASTA format or you may import an alignment. Use the Import Nucleotide FASTA button to import your properly formatted FASTA file. For help with how to format the FASTA file click the FASTA Format Help button. The Sequences tab will display the information about the imported sequence(s). Please check the number of sequences, Sequence IDs (SeqIDs) and length of each sequence to make sure this information is correct. You may also import a nucleotide alignment file in any of the Sequin compatible formats (fasta+gap, Nexus, Phylip). See examples of these formats here.

If the sequences contain a significant number of ambiguous bases near the 5' or 3' end, you may be prompted to trim or remove these sequences from your submission.

Trim Vector Contamination: It is highly recommended that you perform a vector screen on your sequences and trim vector contamination by clicking the Vector Trim Tool button.

Delete Sequences: You can remove sequences from your submission using the Sequence Deletion Tool under the Edit Menu. This tool will assist you in removing any sequences from your file that you need to delete or that do not meet GenBank minimum sequence length requirements.

Sequencing Method

If you are submitting over 500 sequences or your sequences were generated using next-generation sequencing technology, the information in this form is required.

Sequencing Method: Use the check boxes at the top of the form to select the sequencing technology type(s) used to obtain the sequences. Multiple types can be selected, if appropriate. If you used technology that is not listed in the form, please select other and use the free text box to provide the information.

Assembly Program: After selecting the sequencing technology, select the radio button to indicate if your sequences are raw sequence reads or sequence assemblies. If you are submitting assemblies using next-generation sequencing technology, the name of the assembly program and program version or date the assemblies were made are required in the free text boxes. If multiple assembly programs were used, Click on Add More Assembly Programs and complete the provided spreadsheet.

Raw sequence reads from next generation sequencing technologies should not be submitted to GenBank.

Submission Type

If you are submitting more than one sequence, you will be prompted to select the type of submission you are creating. If you select a set, all of the sequences in the set must have the same release date. The following submission types are available in the Virus Wizard:

- Pop set (Population study): a set of sequences that were derived by sequencing the same gene from different isolates of the same organism.
- Phy set (Phylogenetic study): a set of sequences that were derived by sequencing the same gene from different organisms.
- Mut set (Mutation study): a set of sequences that were derived by sequencing multiple mutations of a single gene.
- Batch: related sequences that are not part of a population, mutation, or phylogenetic study. The sequences should be related in some way, such as coming from the same publication or organism.

Virus Wizard Type of Virus

Use this page to select the type of virus sequence(s) you are submitting. There are specific source requirements for certain virus types, including:

- Norovirus, Sapovirus (Caliciviridae)
- Foot-and-mouth disease virus
- Influenza virus
- Rotavirus

If the virus type is not listed or you are submitting sequences from a mixed set of different viruses, select the “Not listed above or mixed set of different viruses" radio button.

Virus Wizard Source Information

Requirements: Each type of virus submission has specific source requirements. Please see the sub-section below for specific requirements for each virus type. In addition to the specific requirements listed below, you will need to provide unique source information (such as unique strain or isolate names/IDs) for each sequence if all of your sequences are from the same gene/region.

How to add source information: There are three ways to add the source information: 1) directly type into this form, 2) import a tab-delimited source table, or 3) automatically populate the form if source information was included in the FASTA definition lines.

You can set the same source qualifier value for all sequences by filling in the top row of boxes and using the appropriate Apply button. Use the Copy from SeqID button to apply the sequence IDs to the qualifier indicated in this table if this information was used as the sequence IDs in the original FASTA file.

Click on the Source Table Help button to open a text dialog with information on making a tab-delimited source table. Click the Export This Table button to export a tab-delimited template file. You must maintain the tab-structure of the table in order to correctly import the data back into the submission wizard. Do not use spaces between the columns.

If you entered all required source information in the FASTA definition lines, minimal input will be necessary on this form.

Errors: Any problems or missing information will be listed on the right side of the form. If you have made any changes on this form, please use the Recheck Errors button to validate the new information. Use the Show only sequences with errors radio button to list only those sequences that did not pass the validation.

Are you unable to pass the Source Information window? If you have not provided some required source information, the issue will be listed in the ***Problems*** column. After fixing any problems, click the Recheck Errors Button to determine if all issues have been fixed. You may display only the entries with problems by selecting the radio button next to Show only sequences with errors.

Do you not see a source qualifier in the table that you want to use in your submission? You may add columns for some commonly added source qualifiers using the buttons below the table. Other optional modifiers can be added to provide additional information using the "Apply/See More Source Information" button or "Import Source Table" button. A window with instructions for creating a source table can be viewed by clicking Source Table Help.

Did you have source information in your FASTA file that is not displayed in this table? This table only displays the required source qualifiers for each type of submission. It does not display all source information. If your FASTA definition lines were correctly formatted, the extra source information you provided in the FASTA definition lines will be imported. You will be able to review this information in the record viewer.

Norovirus, Sapovirus (Caliciviridae) Requirements

Norovirus or Sapovirus (Caliciviridae) submissions must include the following information:

1. organism name
2. isolate
3. collection-date
4. country
5. host or isolation-source

Use the Add isolation-source button if the source of the virus is better described as an isolation source rather than host.

Foot-and-mouth Disease Virus Requirements

Foot-and-mouth disease virus submissions must include the following information:

1. organism name
2. isolate

Use the other fields in the table to provide additional information about the source (country, collection-date, host, etc).

Influenza Virus Requirements

Influenza virus submissions must include the following information:

1. organism name
2. properly formatted strain
3. collection-date
4. host or isolation-source
5. segment
6. country
7. Influenza A viruses also must list the serotype

Use the Add isolation-source button if the source of the virus is better described as an isolation source rather than host. Passage history is optional.

Rotavirus Requirements

Rotavirus submissions must include the following information:

1. organism name
2. isolate
3. collection-date
4. country
5. host or isolation-source

Use the Add isolation-source button if the source of the virus is better described as an isolation source rather than host. Please use the other qualifiers in the table to provide additional information about the source.

Not listed above or mixed set of different viruses Requirements

All other virus sequences must include the following information:

1. organism name
2. isolate
3. country (optional)
4. collection-date (optional)
5. host (optional)

The country, collection-date, and host are optional fields, however we urge you to provide this information for any virus submission. Use the Add isolation-source button if the source of the virus is better described as an isolation source rather than host. You may be prompted at a later date for country, collection-date, and host/isolation-source information if you do not provide it in this table.

Virus Wizard Molecule Information

Use this page to select the molecule type that was isolated and sequenced in your experiment. The topology of the molecule can also be changed in this window. Only set the topology to circular if you are submitting a complete, circular viral genome or segment. Single genes or fragments of viral genomes should not be set to a circular topology.

If you selected mRNA as the molecule type you will be prompted for more information about your samples.

Virus Wizard Annotation

Use the radio buttons to select the option that best describes the sequences. After completing all dialogs for each section, you will be directed to leave the Wizard and transferred to the record viewer. You must do so to complete your submission. However, you cannot return to the Wizard once you have exited.

Note about Influenza Annotation: If you selected Influenza virus as the type of source in a previous dialog, please select "Multiple features per sequence" and follow the dialog instructions for uploading a feature table made using the NCBI Influenza Genome Annotation Tool.

Single coding region across the entire sequence

Select this option if the sequences contain the same, single coding region across the entire length of all of the sequences. Once you have selected this button, a new dialog will appear with text boxes to input the protein name, protein description, gene symbol and comments. Only the protein name is required, other fields are optional. If the coding region is partial, check the appropriate 5' or 3' boxes near the top of the dialog as appropriate.

Single non-coding feature across the entire sequence

Select this option if the sequences contain a single non-coding feature, such as a UTR or LTR, across the entire length of all of the sequences. Once of you have selected this button, a new dialog will appear listing common types of non-coding features. Select the appropriate radio button. If none of the choices listed are appropriate, select “Something else” and a free text box will appear for you to type a description of what the sequences contain.

Multiple features per sequence (coding regions, LTRs, etc.)

Select this option if the sequences contain more than one feature and you know the nucleotide spans of each feature. Once this option is selected, a dialog will open with instructions for importing a five-column, tab-delimited feature table containing all of the feature locations and you will be prompted to exit the wizard and open the record viewer. You may also apply annotation using the Annotate menu options in the record viewer. Alternately, if you imported an alignment you may use Feature Propagate or the Alignment Assistant to add feature annotation to your submission.

If you are submitting Influenza sequences and you selected Influenza virus as the type of virus in a previous dialog, you will be prompted to make and upload a feature table in the dialog that follows. Please follow the instructions for uploading a feature table made using the NCBI Influenza Genome Annotation Tool.

Submission Wizard for Uncultured Samples

Created: May 15, 2012; Updated: January 16, 2014.

Purpose

The Uncultured Sample Submission Wizard is for submitting sequences obtained from an uncultured source/ environmental samples only. Do not use this wizard for sequences from purified bacterial or fungal strains. This wizard will guide you in providing all of the necessary source information for uncultured samples and will provide assistance and direction with feature annotation. Examples of source information are provided.

Wizard Import Nucleotide Sequences

Requirements: The Uncultured Sample Sequence Submission requirements are listed in the Sequences tab of the Wizard Import Nucleotide Sequences dialog box.

Sequence Format: You may import your sequences in FASTA format or you may import an alignment. Use the Import Nucleotide FASTA button to import your properly formatted FASTA file. For help with how to format the FASTA file click the FASTA Format Help button. The Sequences tab will display the information about the imported sequence(s). Please check the number of sequences, Sequence IDs (SeqIDs) and length of each sequence to make sure this information is correct. You may also import a nucleotide alignment file in any of the Sequin compatible formats (fasta+gap, Nexus, Phylip). See examples of these formats here.

If the sequences contain a significant number of ambiguous bases near the 5' or 3' end, you may be prompted to trim or remove these sequences from your submission.

Trim Vector Contamination: It is highly recommended that you perform a vector screen on your sequences and trim vector contamination by clicking the Vector Trim Tool button.

Delete Sequences: You can remove sequences from your submission using the Sequence Deletion Tool under the Edit Menu. This tool will assist you in removing any sequences from your file that you need to delete or that do not meet GenBank minimum sequence length requirements.

Sequencing Method

If you are submitting over 500 sequences or your sequences were generated using next-generation sequencing technology, the information in this form is required.

Sequencing Method: Use the check boxes at the top of the form to select the sequencing technology type(s) used to obtain the sequences. Multiple types can be selected, if appropriate. If you used technology that is not listed in the form, please select other and use the free text box to provide the information.

Assembly Program: After selecting the sequencing technology, select the radio button to indicate if your sequences are raw sequence reads or sequence assemblies. If you are submitting assemblies using next-generation sequencing technology, the name of the assembly program and program version or date the assemblies were made are required in the free text boxes. If multiple assembly programs were used, Click on Add More Assembly Programs and complete the provided spreadsheet.

Raw sequence reads from next generation sequencing technologies should not be submitted to GenBank.

Submission Type

If you imported a nucleotide FASTA file and you are submitting more than one sequence, you will be prompted to select the type of submission you are creating. This dialog will not appear if you imported an alignment. If you

select a set, all of the sequences in the set must have the same release date. The following submission types are available in the uncultured Wizard:

- Environmental Set: a set of sequences that were derived by sequencing the same gene from a population of unclassified or unknown organisms.
- Batch: related sequences that are not part of a population, mutation, or phylogenetic study. The sequences should be related in some way, such as coming from the same publication or organism.

Uncultured Sample Wizard Source Information

Requirements: All sequences must have an organism name, isolation-source or host, and unique clone name. Optional modifiers can be added to provide additional information, if known. The organism name should not contain the entire lineage information. Please review examples of uncultured sample organism names.

The clone name is used for both traditional clones and PCR product sample IDs. The clone names must be unique within the submission. In addition, information about the environment from which the sequences were isolated should be supplied within the isolation-source field. If the source organism was isolated from within a host organism, this should be supplied in the host field.

How to add source information: There are three ways to add the source information: 1) directly type into this form, 2) import a tab-delimited source table, or 3) automatically populate the form if source information was included in the FASTA definition lines.

You can set the same source qualifier value for all sequences by filling in the top row of boxes and using the appropriate Apply button. Use the Copy from SeqID button to apply the sequence IDs to the qualifier indicated in this table if this information was used as the sequence IDs in the original FASTA file.

Click on the Source Table Help button to open a text dialog with information on making a tab-delimited source table.

If you entered all required source information in the FASTA definition lines, minimal input will be necessary on this form.

Errors: Any problems or missing information will be listed on the right side of the form. If you have made any changes on this form, please use the Recheck Errors button to validate the new information. Use the Show only sequences with errors radio button to list only those sequences that did not pass the validation.

Are you unable to pass the Source Information window? If you have not provided some required source information, the issue will be listed in the ***Problems*** column. After fixing any problems, click the Recheck Errors Button to determine if all issues have been fixed. You may display only the entries with problems by selecting the radio button next to Show only sequences with errors.

Do you not see a source qualifier in the table that you want to use in your submission? You may add columns for some commonly added source qualifiers using the buttons below the table. Other optional modifiers can be added to provide additional information using the "Apply/See More Source Information" button or "Import Source Table" button. A window with instructions for creating a source table can be viewed by clicking Source Table Help.

Did you have source information in your FASTA file that is not displayed in this table? This table only displays the required source qualifiers for each type of submission. It does not display all source information. If your FASTA definition lines were correctly formatted, the extra source information you provided in the FASTA definition lines will be imported. You will be able to review this information in the record viewer.

Uncultured Sample Wizard Primer Type

Use this page to select the type of primers used in PCR amplification of the samples.

Select “universal primers” if the primers amplify DNA from a broad range of organisms.

Select “species-specific primers” if primers amplify DNA from a single species.

Uncultured Sample Annotation

Use the radio buttons to select the option that best describes the sequences. After completing all dialogs for each section, you will be directed to leave the Wizard and transferred to the record viewer. You must do so to complete your submission. However, you cannot return to the Wizard once you have exited.

Single rRNA, ITS, or IGS

Select this option if the sequences contain a single ribosomal RNA, one internal transcribed spacer, or one intergenic spacer across the entire sequence. Once you have selected this button, a new dialog will appear with radio buttons to select the type of organism from which the sequences were derived. Once the organism type is selected, a dialog listing common types of rRNA, ITS, or IGS will appear. Select the appropriate radio button. If none of the choices are appropriate, select Something else and type in a description of what the sequences contain.

Multiple rRNA, ITS, or IGS regions where spans are unknown

Select this option if the sequences contain more than one ribosomal RNA, internal transcribed spacer, and/or intergenic spacer and you are uncertain of the nucleotide locations of each feature. Once you have selected this button, a new dialog will appear with radio buttons to select the type of organism from which the sequences were derived. Once the organism type is selected, a dialog listing common types of rRNA, ITS, and IGS will appear. Select the appropriate checkboxes. If none of the choices are appropriate, select Something else and type in a description of what the sequences contain.

Multiple rRNA, ITS, or IGS where spans are known

Select this option if the sequences contain more than one ribosomal RNA, internal transcribed spacer, and/or intergenic spacer and you know the nucleotide spans of each feature. Once this option is selected, you will be prompted to exit the wizard and the record viewer will open. A text dialog will open with instructions for importing a five-column, tab-delimited table containing the feature locations. You may also apply annotation using the Annotate menu options in the record viewer. Alternately, if you imported an alignment you may use Feature Propagate or the Alignment Assistant to add feature annotation to your submission.

Intergenic spacer (not rRNA-IGS)

Select this option if your sequences contain intergenic spacer that is not ribosomal intergenic spacer. The dialogs that follow will prompt you for more information about the flanking genes and if these genes are present in the sequences. Please see more documentation here.

Coding Region (CDS)

Select this option if the sequences encode the same, single protein across the entire length of the sequence. Once you have selected this button, a new dialog will appear with text boxes to input the protein name, protein description, gene symbol, and comments. Only the protein name is required, other fields are optional. If the coding region is partial, check the appropriate 5’ or 3’ boxes near the top of the dialog as appropriate.

Something else, multiple features

Select this option if the sequences do not contain one of the feature types listed in the annotation dialog or you wish to apply annotation in the record viewer. Upon selecting this option, you will be prompted to exit the wizard and the record viewer will open. A text dialog will open with instructions for importing a five-column, tab-delimited table containing the feature locations. You may also apply annotation using the Annotate menu options in the record viewer. Alternately, if you imported an alignment you may use Feature Propagate or the Alignment Assistant to add feature annotation to your submission.

Wizard rRNA Chimera Checking

If you selected the annotation of 16S ribosomal RNA sequences from bacteria or archaea, an additional dialog will appear asking if the sequences have been screened with a chimera check program. If you have used a chimera check program, please provide the name and version if applicable. Note that BLAST is not a chimera check program. If you have screened the sequences for chimeras, please be sure to remove all suspected chimeric sequences before submission.

Submission Wizard for rRNA-ITS-IGS Sequences

Created: May 15, 2012; Updated: January 16, 2014.

Purpose

The rRNA-ITS-IGS Sequence Submission Wizard is for submitting rRNA, internal transcribed spacer, and rRNA-intergenic spacer sequences obtained from:

- pure, cultured bacterial samples
- pure, cultured archaeal samples
- pure, cultured or vouchered fungal samples
- plant samples
- animal samples

Do not use this wizard for sequences from an uncultured source. This wizard will guide you in providing all of the necessary source information for rRNA/ITS/IGS samples and will provide assistance and direction with feature annotation. Examples of source information are provided.

Wizard Import Nucleotide Sequences

Requirements: The rRNA-ITS-IGS Sequence Submission requirements are listed in the Sequences tab of the Wizard Import Nucleotide Sequences dialog box.

Sequence Format: You may import your sequences in FASTA format or you may import an alignment. Use the Import Nucleotide FASTA button to import your properly formatted FASTA file. For help with how to format the FASTA file click the FASTA Format Help button. The Sequences tab will display the information about the imported sequence(s). Please check the number of sequences, Sequence IDs (SeqIDs) and length of each sequence to make sure this information is correct. You may also import a nucleotide alignment file in any of the Sequin compatible formats (fasta+gap, Nexus, Phylip). See examples of these formats here.

If the sequences contain a significant number of ambiguous bases near the 5' or 3' end, you may be prompted to trim or remove these sequences from your submission.

Trim Vector Contamination: It is highly recommended that you perform a vector screen on your sequences and trim vector contamination by clicking the Vector Trim Tool button.

Delete Sequences: You can remove sequences from your submission using the Sequence Deletion Tool under the Edit Menu. This tool will assist you in removing any sequences from your file that you need to delete or that do not meet GenBank minimum sequence length requirements.

Sequencing Method

If you are submitting over 500 sequences or your sequences were generated using next-generation sequencing technology, the information in this form is required.

Sequencing Method: Use the check boxes at the top of the form to select the sequencing technology type(s) used to obtain the sequences. Multiple types can be selected, if appropriate. If you used technology that is not listed in the form, please select other and use the free text box to provide the information.

Assembly Program: After selecting the sequencing technology, select the radio button to indicate if your sequences are raw sequence reads or sequence assemblies. If you are submitting assemblies using next-generation sequencing technology, the name of the assembly program and program version or date the

assemblies were made are required in the free text boxes. If multiple assembly programs were used, Click on Add More Assembly Programs and complete the provided spreadsheet.

Raw sequence reads from next generation sequencing technologies should not be submitted to GenBank.

Submission Type

If you are submitting more than one sequence, you will be prompted to select the type of submission you are creating. If you select a set, all of the sequences in the set must have the same release date. The following submission types are available in the rRNA-ITS-IGS Wizard:

- Pop set (Population study): a set of sequences that were derived by sequencing the same gene from different isolates of the same organism.
- Phy set (Phylogenetic study): a set of sequences that were derived by sequencing the same gene from different organisms.
- Mut set (Mutation study): a set of sequences that were derived by sequencing multiple mutations of a single gene.
- Batch: related sequences that are not part of a population, mutation, or phylogenetic study. The sequences should be related in some way, such as coming from the same publication or organism. This option will not be available if you imported an alignment.

rRNA-ITS-IGS Wizard Type of Source

Use this page to select the type of organism from which your sequences were obtained.

rRNA-ITS-IGS Wizard Source Information

Requirements: All sequences must have an organism name. Sequences from bacteria and archaea must have unique strain names. Fungal sequences must have unique strain names, unique specimen-vouchers, or a combination of unique organism names, specimen-vouchers, and isolate codes. Sequences from all other organisms must have either unique organism names or a combination of unique organism names and isolate names.

How to add source information: There are three ways to add the source information: 1) directly type into this form, 2) import a tab-delimited source table, or 3) automatically populate the form if source information was included in the FASTA definition lines.

You can set the same source qualifier value for all sequences by filling in the top row of boxes and using the appropriate Apply button. Use the Copy from SeqID button to apply the sequence IDs to the qualifier indicated in this table if this information was used as the sequence IDs in the original FASTA file.

Click on the Source Table Help button to open a text dialog with information on making a tab-delimited source table.

If you entered all required source information in the FASTA definition lines, minimal input will be necessary on this form.

Errors: Any problems or missing information will be listed on the right side of the form. If you have made any changes on this form, please use the Recheck Errors button to validate the new information. Use the Show only sequences with errors radio button to list only those sequences that did not pass the validation.

Are you unable to pass the Source Information window? If you have not provided some required source information, the issue will be listed in the ***Problems*** column. After fixing any problems, click the Recheck

Errors Button to determine if all issues have been fixed. You may display only the entries with problems by selecting the radio button next to Show only sequences with errors.

Do you not see a source qualifier in the table that you want to use in your submission? You may add columns for some commonly added source qualifiers using the buttons below the table. Other optional modifiers can be added to provide additional information using the "Apply/See More Source Information" button or "Import Source Table" button. A window with instructions for creating a source table can be viewed by clicking Source Table Help.

Did you have source information in your FASTA file that is not displayed in this table? This table only displays the required source qualifiers for each type of submission. It does not display all source information. If your FASTA definition lines were correctly formatted, the extra source information you provided in the FASTA definition lines will be imported. You will be able to review this information in the record viewer.

rRNA-ITS-IGS Wizard Genome

This dialog will only appear if you selected Cultured Fungus, Vouchered Fungus, or Something else in the rRNA-ITS-IGS Wizard Type of Source form. Use this page to select the location from which your sequences are derived. If your sequences are derived from nuclear genomic DNA, select nuclear. If your sequences are derived from an alternative location (i.e. organelle genome), select the appropriate location in this dialog.

rRNA-ITS-IGS Wizard Annotation

Use the radio buttons and checkboxes to select the option that best describes the sequences. After completing all dialogs for each section, you will be directed to leave the wizard and transferred to the record viewer. You must exit the wizard to complete your submission. However, you cannot return to the Wizard once you have exited. Note that the availability of some options in the annotation dialogs depends on your previous selections in the wizard.

Single rRNA or IGS

You will only have this option if you selected Cultured Bacteria or Archaea as the Type of Source.

Select this option if the sequences only contain a single ribosomal RNA or one intergenic spacer across the entire sequence. Once you have selected this button, a new dialog listing common types of rRNA and IGS will appear. You will only be able to select one feature. Select the appropriate radio button. If none of the choices are appropriate, select Something else and type in a description of what the sequences contain.

Multiple rRNA or IGS where spans are unknown

You will only have this option if you selected Cultured Bacteria or Archaea as the Type of Source.

Select this option if the sequences contain more than one feature (for example 16S ribosomal RNA and intergenic spacer) and you are uncertain about the nucleotide locations of each feature. Once you have selected this button, a new dialog listing common types of rRNA and IGS will appear. Select the appropriate checkboxes. If none of the choices are appropriate, select Something else and type in a description of what the sequences contain.

Multiple rRNA or IGS where spans are known

You will only have this option if you selected Cultured Bacteria or Archaea as the Type of Source.

Select this option if the sequences contain more than one feature (for example 16S ribosomal RNA and intergenic spacer) and you know the nucleotide spaces of each feature. Once this option is selected, you will be

prompted to exit the wizard and the record viewer will open. A text dialog will open with instructions for importing a five-column, tab-delimited table containing the feature locations. You may also apply annotation using the Annotate menu options in the record viewer. Alternately, if you imported an alignment you may use Feature Propagate or the Alignment Assistant to add feature annotation to your submission.

Single rRNA or ITS

You will only have this option if you selected Cultured Fungus, Vouchered Fungus, or Something else as the Type of Source.

Select this option if the sequences only contain a single ribosomal RNA or one internal transcribed spacer across the entire sequence. Once you have selected this button, a new dialog listing common types of rRNA and ITS will appear. You will only be able to select one feature. Select the appropriate radio button. If none of the choices are appropriate, select Something else and type in a description of what the sequences contain.

Multiple rRNA or ITS where spans are unknown

You will only have this option if you selected Cultured Fungus, Vouchered Fungus, or Something else as the Type of Source.

Select this option if the sequences contain more than one feature (for example 18S ribosomal RNA and internal transcribed spacer 1) and you are uncertain about the nucleotide locations of each feature. Once you have selected this button, a new dialog listing common types of rRNA and ITS will appear. Select the appropriate checkboxes. If none of the choices are appropriate, select Something else and type in a description of what the sequences contain.

Multiple rRNA or IGS where spans are known

You will only have this option if you selected Cultured Fungus, Vouchered Fungus, or Something else as the Type of Source.

Select this option if the sequences contain more than one feature (for example 18S ribosomal RNA and internal transcribed spacer 1) and you know the nucleotide spaces of each feature. Once this option is selected, you will be prompted to exit the wizard and the record viewer will open. A text dialog will open with instructions for importing a five-column, tab-delimited table containing the feature locations. You may also apply annotation using the Annotate menu options in the record viewer. Alternately, if you imported an alignment you may use Feature Propagate or the Alignment Assistant to add feature annotation to your submission.

Something else

Select this option if the sequences do not contain one of the feature types listed in the annotation dialog or you wish to apply annotation in the record viewer. Upon selecting this option, you will be prompted to exit the wizard and the record viewer will open. A text dialog will open with instructions for importing a five-column, tab-delimited table containing the feature locations. You may also apply annotation using the Annotate menu options in the record viewer. Alternately, if you imported an alignment you may use Feature Propagate or the Alignment Assistant to add feature annotation to your submission.

Wizard rRNA Chimera Checking

If you selected the annotation of 16S ribosomal RNA sequences from bacteria or archaea, an additional dialog will appear asking if the sequences have been screened with a chimera check program. If you have used a chimera check program, please provide the name and version if applicable. Note that BLAST is not a chimera check program. If you have screened the sequences for chimeras, please be sure to remove all suspected chimeric sequences before submission.

Submission Wizard for Intergenic Spacer (IGS) Sequences

Created: May 15, 2012; Updated: January 16, 2014.

Purpose

The Intergenic Spacer (IGS) Sequence Submission Wizard is for submitting non-rRNA intergenic spacer sequences from pure, cultured fungal strains or vouchered fungal samples, plant, or animal sequences. This wizard will guide you in providing all of the necessary source information for these types of samples and will provide assistance and direction with feature annotation. Examples of source information are provided.

Wizard Import Nucleotide Sequences

Requirements: The Intergenic Spacer (IGS) Sequence Submission requirements are listed in the Sequences tab of the Wizard Import Nucleotide Sequences dialog box.

Sequence Format: You may import your sequences in FASTA format or you may import an alignment. Use the Import Nucleotide FASTA button to import your properly formatted FASTA file. For help with how to format the FASTA file click the FASTA Format Help button. The Sequences tab will display the information about the imported sequence(s). Please check the number of sequences, Sequence IDs (SeqIDs) and length of each sequence to make sure this information is correct. You may also import a nucleotide alignment file in any of the Sequin compatible formats (fasta+gap, Nexus, Phylip). See examples of these formats here.

If the sequences contain a significant number of ambiguous bases near the 5' or 3' end, you may be prompted to trim or remove these sequences from your submission.

Trim Vector Contamination: It is highly recommended that you perform a vector screen on your sequences and trim vector contamination by clicking the Vector Trim Tool button.

Delete Sequences: You can remove sequences from your submission using the Sequence Deletion Tool under the Edit Menu. This tool will assist you in removing any sequences from your file that you need to delete or that do not meet GenBank minimum sequence length requirements.

Sequencing Method

If you are submitting over 500 sequences or your sequences were generated using next-generation sequencing technology, the information in this form is required.

Sequencing Method: Use the check boxes at the top of the form to select the sequencing technology type(s) used to obtain the sequences. Multiple types can be selected, if appropriate. If you used technology that is not listed in the form, please select other and use the free text box to provide the information.

Assembly Program: After selecting the sequencing technology, select the radio button to indicate if your sequences are raw sequence reads or sequence assemblies. If you are submitting assemblies using next-generation sequencing technology, the name of the assembly program and program version or date the assemblies were made are required in the free text boxes. If multiple assembly programs were used, Click on Add More Assembly Programs and complete the provided spreadsheet.

Raw sequence reads from next generation sequencing technologies should not be submitted to GenBank.

Submission Type

If you are submitting more than one sequence, you will be prompted to select the type of submission you are creating. If you select a set, all of the sequences in the set must have the same release date. The following submission types are available in the IGS Wizard:

- Pop set (Population study): a set of sequences that were derived by sequencing the same gene from different isolates of the same organism.
- Phy set (Phylogenetic study): a set of sequences that were derived by sequencing the same gene from different organisms.
- Mut set (Mutation study): a set of sequences that were derived by sequencing multiple mutations of a single gene.
- Batch: related sequences that are not part of a population, mutation, or phylogenetic study. The sequences should be related in some way, such as coming from the same publication or organism. This option will not be available if you imported an alignment.

IGS Wizard Type of Source

Use this page to select the type of organism from which your sequences were obtained.

IGS Wizard Source Information

Requirements: All sequences must have an organism name. Sequences obtained from cultured fungi must have unique strain names. Sequences from vouchered fungal samples must have specimen-vouchers and a unique combination of organism names, specimen-vouchers, and isolate codes. Sequences from all other organisms must have unique organism names or a combination or unique organism names, isolate names, and/or specimen-vouchers.

How to add source information: There are three ways to add the source information: 1) directly type into this form, 2) import a tab-delimited source table, or 3) automatically populate the form if source information was included in the FASTA definition lines.

You can set the same source qualifier value for all sequences by filling in the top row of boxes and using the appropriate Apply button. Use the Copy from SeqID button to apply the sequence IDs to the qualifier indicated in this table if this information was used as the sequence IDs in the original FASTA file.

Click on the Source Table Help button to open a text dialog with information on making a tab-delimited source table.

If you entered all required source information in the FASTA definition lines, minimal input will be necessary on this form.

Errors: Any problems or missing information will be listed on the right side of the form. If you have made any changes on this form, please use the Recheck Errors button to validate the new information. Use the Show only sequences with errors radio button to list only those sequences that did not pass the validation.

Are you unable to pass the Source Information window? If you have not provided some required source information, the issue will be listed in the ***Problems*** column. After fixing any problems, click the Recheck Errors Button to determine if all issues have been fixed. You may display only the entries with problems by selecting the radio button next to Show only sequences with errors.

Do you not see a source qualifier in the table that you want to use in your submission? You may add columns for some commonly added source qualifiers using the buttons below the table. Other optional modifiers can be

added to provide additional information using the "Apply/See More Source Information" button or "Import Source Table" button. A window with instructions for creating a source table can be viewed by clicking Source Table Help.

Did you have source information in your FASTA file that is not displayed in this table? This table only displays the required source qualifiers for each type of submission. It does not display all source information. If your FASTA definition lines were correctly formatted, the extra source information you provided in the FASTA definition lines will be imported. You will be able to review this information in the record viewer.

IGS Wizard Genome

Use this page to select the location from which your sequences were derived. If your sequences were derived from nuclear genomic DNA, select nuclear. If your sequences are derived from an alternative location (i.e. organelle genome), select the appropriate location in this dialog.

IGS Wizard Annotation

Use the radio buttons to select the option that best describes the sequences. After completing all dialogs for each section, you will be directed to leave the wizard and transferred to the record viewer. You must exit the Wizard to complete your submission. However, you cannot return to the Wizard once you have exited.

Intergenic spacer only

Select this option if your sequences only contain intergenic spacer. If your sequences contain part of the flanking gene(s), do not select this option. Once this option is selected, a new dialog will prompt you for the gene symbols of the genes flanking the intergenic spacer at the 5' and 3' ends. Type the gene symbol for the gene flanking the spacer at the 5' end in the text box below 5' gene symbol. Type the gene symbol for the gene flanking the spacer at the 3' end in the text box below 3' gene symbol. You will then need to select the radio button to indicate if your sequences contain the partial or complete intergenic spacer at the 5' and 3' ends.

Intergenic spacer and other features (gene, tRNA) where spans are unknown

Select this option if your sequences contain intergenic spacer and part of the flanking gene(s) and you do not know the exact nucleotide locations of the features in your sequences. Once this option is selected, a new dialog will prompt you for the type of features flanking the spacer. Select the appropriate feature for the 5' end and 3' end. If you select tRNA, you will be prompted to select the name of the tRNA from a list of tRNAs. If you select protein coding gene, you will be provided text boxes to type in the protein name and gene symbol. After you have provided feature information, use the radio buttons to indicate if your sequences contain part of the features at the 5' end and 3' end. If the feature is not present at one or both ends, you will need to indicate if the intergenic spacer is partial or complete.

Intergenic spacers and other features (gene, tRNA) where spans are known

Select this option if the sequences contain intergenic spacer and other features (coding region, tRNA, etc.) and you know the nucleotide locations of the features in your sequences. Once this option is selected, you will be prompted to exit the wizard and the record viewer will open. A text dialog will open with instructions for importing a five-column, tab-delimited table containing the feature locations. You may also apply annotation using the Annotate menu options in the record viewer. Alternately, if you imported an alignment you may use Feature Propagate or the Alignment Assistant to add feature annotation to your submission.

Something else

Select this option if the sequences do not contain one of the feature types listed in the annotation dialog or you wish to apply annotation in the record viewer. Upon selecting this option, you will be prompted to exit the wizard and the record viewer will open. A text dialog will open with instructions for importing a five-column, tab-delimited table containing the feature locations. You may also apply annotation using the Annotate menu options in the record viewer. Alternately, if you imported an alignment you may use Feature Propagate or the Alignment Assistant to add feature annotation to your submission.

Submission Wizard for Microsatellite sequences

Created: May 15, 2012; Updated: January 16, 2014.

Purpose

The Microsatellite Sequence Submission Wizard is for submitting microsatellite sequence submissions only. This wizard will guide you in providing all of the necessary source information for microsatellite sequences and will provide assistance and direction with feature annotation. Examples of source information are provided.

Wizard Import Nucleotide Sequences

Requirements: The Microsatellite Submission requirements are listed in the Sequences tab of the Wizard Import Nucleotide Sequences dialog box.

Sequence Format: Use the Import Nucleotide FASTA button to import your properly formatted FASTA file. For help with how to format the FASTA file click the FASTA Format Help button. The Sequences tab will display the information about the imported sequence(s). Please check the number of sequences, Sequence IDs (SeqIDs) and length of each sequence to make sure this information is correct.

If the sequences contain a significant number of ambiguous bases near the 5' or 3' end, you may be prompted to trim or remove these sequences from your submission.

Trim Vector Contamination: It is highly recommended that you perform a vector screen on your sequences and trim vector contamination by clicking the Vector Trim Tool button.

Delete Sequences: You can remove sequences from your submission using the Sequence Deletion Tool under the Edit Menu. This tool will assist you in removing any sequences from your file that you need to delete or that do not meet GenBank minimum sequence length requirements.

Sequencing Method

If you are submitting over 500 sequences or your sequences were generated using next-generation sequencing technology, the information in this form is required.

Sequencing Method: Use the check boxes at the top of the form to select the sequencing technology type(s) used to obtain the sequences. Multiple types can be selected, if appropriate. If you used technology that is not listed in the form, please select other and use the free text box to provide the information.

Assembly Program: After selecting the sequencing technology, select the radio button to indicate if your sequences are raw sequence reads or sequence assemblies. If you are submitting assemblies using next-generation sequencing technology, the name of the assembly program and program version or date the assemblies were made are required in the free text boxes. If multiple assembly programs were used, Click on Add More Assembly Programs and complete the provided spreadsheet.

Raw sequence reads from next generation sequencing technologies should not be submitted to GenBank.

Microsatellite Wizard Molecule Type

Use this page to select the molecule type that was isolated and sequenced in your experiment. The default selection is nuclear genomic DNA. However, if you did not isolate genomic DNA from the nucleus, you should select no to the first question and then select the appropriate genome location and molecule type.

Microsatellite Wizard Annotation

Use this page to select what type of annotation you are planning to add to your sequence submission.

To apply a single repeat_region across the length of all of your sequences, select the Apply 1 microsatellite across entire sequence(s) radio button. If you have specific information about the repeat, such as the sequence of the repeat (rpt_unit_seq) or the nucleotide location of one repeat unit (rpt_unit_range), select the checkboxes next to those options. The rpt_unit_seq and rpt_unit_range information is optional.

To apply more complex annotation, such as multiple repeat_regions per sequence or repeat_regions with specific nucleotide spans, click Apply multiple microsatellites. After you provide the required source information, you will be prompted to add this information in the record viewer. A text dialog will open with instructions for importing a five-column, tab-delimited table containing the feature locations. You may also apply annotation using the Annotate menu options in the record viewer.

Microsatellite Wizard Information

Requirements: All sequences must have an organism name and a microsatellite name or clone name. After you provide all required information you will be prompted to exit the wizard and the record viewer will open.

How to add information to this table: There are three ways to add the information: 1) directly type into this form, 2) import a tab-delimited table, or 3) automatically populate the source information into the form if source information was included in the FASTA definition lines.

You can set the same qualifier value for all sequences by filling in the top row of boxes and using the appropriate Apply button. Use the Copy from SeqID button to apply the sequence IDs to the qualifier indicated in this table if this information was used as the sequence IDs in the original FASTA file.

Click on the Source Table Help button to open a text dialog with information on making a tab-delimited table. Click the Export This Table button to export a tab-delimited template file. You must maintain the tab-structure of the table in order to correctly import the data back into the submission wizard. Do not use spaces between the columns.

Errors: Any problems or missing information will be listed on the right side of the form. If you have made any changes on this form, please use the Recheck Errors button to validate the new information. Use the Show only sequences with errors radio button to list only those sequences that did not pass the validation.

Are you unable to pass the Information window? If you have not provided some required source information, the issue will be listed in the ***Problems*** column. After fixing any problems, click the Recheck Errors Button to determine if all issues have been fixed. You may display only the entries with problems by selecting the radio button next to Show only sequences with errors.

Do you not see a source qualifier in the table that you want to use in your submission? You may add columns for some commonly added source qualifiers using the buttons below the table. Other optional modifiers can be added to provide additional information using the "Apply/See More Source Information" button or "Import Source Table" button. A window with instructions for creating a source table can be viewed by clicking Source Table Help.

Did you have source information in your FASTA file that is not displayed in this table? This table only displays the required source qualifiers for each type of submission. It does not display all source information. If your FASTA definition lines were correctly formatted, the extra source information you provided in the FASTA definition lines will be imported. You will be able to review this information in the record viewer.

Submission Wizard for D-loops and Control Regions

Created: May 15, 2012; Updated: January 16, 2014.

Purpose

The D-loop and Control Region Submission Wizard is for submitting mitochondrial D-loop and Control Region sequence submissions only. This wizard will guide you in providing all of the necessary source information for these sequences and will provide assistance and direction with feature annotation. Examples of source information are provided.

Wizard Import Nucleotide Sequences

Requirements: The D-loop and Control Region Submission requirements are listed in the Sequences tab of the Wizard Import Nucleotide Sequences dialog box.

Sequence Format: You may import your sequences in FASTA format or you may import an alignment. Use the Import Nucleotide FASTA button to import your properly formatted FASTA file. For help with how to format the FASTA file click the FASTA Format Help button. The Sequences tab will display the information about the imported sequence(s). Please check the number of sequences, Sequence IDs (SeqIDs) and length of each sequence to make sure this information is correct. You may also import a nucleotide alignment file in any of the Sequin compatible formats (fasta+gap, Nexus, Phylip). See examples of these formats here.

If the sequences contain a significant number of ambiguous bases near the 5' or 3' end, you may be prompted to trim or remove these sequences from your submission.

Trim Vector Contamination: It is highly recommended that you perform a vector screen on your sequences and trim vector contamination by clicking the Vector Trim Tool button.

Delete Sequences: You can remove sequences from your submission using the Sequence Deletion Tool under the Edit Menu. This tool will assist you in removing any sequences from your file that you need to delete or that do not meet GenBank minimum sequence length requirements.

Sequencing Method

If you are submitting over 500 sequences or your sequences were generated using next-generation sequencing technology, the information in this form is required.

Sequencing Method: Use the check boxes at the top of the form to select the sequencing technology type(s) used to obtain the sequences. Multiple types can be selected, if appropriate. If you used technology that is not listed in the form, please select other and use the free text box to provide the information.

Assembly Program: After selecting the sequencing technology, select the radio button to indicate if your sequences are raw sequence reads or sequence assemblies. If you are submitting assemblies using next-generation sequencing technology, the name of the assembly program and program version or date the assemblies were made are required in the free text boxes. If multiple assembly programs were used, Click on Add More Assembly Programs and complete the provided spreadsheet.

Raw sequence reads from next generation sequencing technologies should not be submitted to GenBank.

Submission Type

If you are submitting more than one sequence, you will be prompted to select the type of submission you are creating. If you select a set, all of the sequences in the set must have the same release date. The following submission types are available in the D-loop Wizard:

- Pop set (Population study): a set of sequences that were derived by sequencing the same gene from different isolates of the same organism.
- Phy set (Phylogenetic study): a set of sequences that were derived by sequencing the same gene from different organisms.
- Mut set (Mutation study): a set of sequences that were derived by sequencing multiple mutations of a single gene.
- Batch: related sequences that are not part of a population, mutation, or phylogenetic study. The sequences should be related in some way, such as coming from the same publication or organism. This option will not be available if you imported an alignment.

D-Loop Wizard Source Information

Requirements: All sequences must have an organism name and unique source information. Use the buttons below the table to add a column to the table, as appropriate. If you are submitting sequences from the same type of organism, you will need to provide one of the following: isolate codes, haplotypes, specimen-voucher, breed, or cultivars. There must be unique source information for each sequence to pass this dialog.

How to add source information: There are three ways to add the source information: 1) directly type into this form, 2) import a tab-delimited source table, or 3) automatically populate the form if source information was included in the FASTA definition lines.

You can set the same source qualifier value for all sequences by filling in the top row of boxes and using the appropriate Apply button. Use the Copy from SeqID button to apply the sequence IDs to the qualifier indicated in this table if this information was used as the sequence IDs in the original FASTA file.

Click on the Source Table Help button to open a text dialog with information on making a tab-delimited source table.

If you entered all required source information in the FASTA definition lines, minimal input will be necessary on this form.

Errors: Any problems or missing information will be listed on the right side of the form. If you have made any changes on this form, please use the Recheck Errors button to validate the new information. Use the Show only sequences with errors radio button to list only those sequences that did not pass the validation.

Are you unable to pass the Source Information window? If you have not provided some required source information, the issue will be listed in the ***Problems*** column. After fixing any problems, click the Recheck Errors Button to determine if all issues have been fixed. You may display only the entries with problems by selecting the radio button next to Show only sequences with errors.

Do you not see a source qualifier in the table that you want to use in your submission? You may add columns for some commonly added source qualifiers using the buttons below the table. Other optional modifiers can be added to provide additional information using the “Apply/See More Source Information” button or “Import Source Table” button. A window with instructions for creating a source table can be viewed by clicking Source Table Help.

Did you have source information in your FASTA file that is not displayed in this table? This table only displays the required source qualifiers for each type of submission. It does not display all source information. If your FASTA definition lines were correctly formatted, the extra source information you provided in the FASTA definition lines will be imported. You will be able to review this information in the record viewer.

D-Loop Wizard Annotation

Use this dialog to select the feature(s) in your sequences. If your sequences contain more than the D-loop/Control Region, select D-loop/Control Region and other features (tRNA/rRNA).

D-Loop Wizard Features

You will be prompted for further information if you selected D-loop/Control Region and other features (tRNA/rRNA). If you do not know the nucleotide spans of the features in your sequences, select No in this dialog and then select the number of other features in your sequences. The D-Loop wizard will only assist with annotating up to four features. If you do know the nucleotide spans of the features in your sequences, select yes in this dialog. A text dialog will open with instructions for importing a five-column, tab-delimited table containing the feature locations. You may also apply annotation using the Annotate menu options in the record viewer. Alternately, if you imported an alignment you may use Feature Propagate or the Alignment Assistant to add feature annotation to your submission.

D-Loop Wizard Feature Annotation

If you indicated that you do not know the nucleotide spans for the features in your sequences, this dialog will collect information for what your sequences contain and create a misc_feature that spans the length of the sequences in your submission. Select the features in your sequences starting at the 5' end as feature 1. If your sequences contain a tRNA or rRNA, click the radio button next to tRNA or rRNA and use the pull-down list to select the appropriate tRNA or rRNA. You may only select D-loop or Control Region once.

Glossary

Created: April 6, 2011.

Annotate/Annotation

The standard definition of the word annotate is "The act or process of furnishing [a literary work with] critical commentary or explanatory notes" (American Heritage Dictionary).

When applied to nucleic acid sequence, the act of annotation identifies the location and the type of feature (e.g. exon, intron, gene, etc.) or provides additional information in the form of a feature modifier (e.g. frequency, function, product, etc.) for a specific region of sequence.

Another way of saying this is that when you annotate a specific region of sequence, you are providing notes that pinpoint the location and describe the biological significance of that region to anyone else looking at the sequence.

Clone

A clone is the identifier (ID number) of a DNA fragment that passively replicates in a host organism after being joined to cloning vector.

Collection Code

A collection code is an identifier that your institution gives to the particular collection from which a specimen came.

Controlled Vocabulary

A controlled vocabulary is a set of predefined, authorized terms that are selected by a specific institution for indexing and retrieval of information.

EST

Expressed Sequence Tags are short (300-500 bp) single reads from mRNA (cDNA) that are usually produced in large numbers. They represent a snapshot of what is expressed in a given tissue, and/or at a given developmental stage. They also represent tags (some coding, others not) of expression for a given cDNA library.

Feature

A feature is a single word or abbreviation that identifies a functional group found on nucleic acid sequence (e.g. exon, intron, gene, coding region, etc). The act of annotating a feature onto a specific region of sequence allows the submitter to provide important information about that region in the sequence record.

The assignment of features to a record permits a user to quickly find/retrieve features, so all the information about a specific feature in a record can be found by using the feature name as a search term. For more information about features as well as a list of features and their definitions, see the DDBJ/EMBL Feature Table Definition page, specifically section 3.2 "Feature Keys".

Isolation_Source

An isolation source is the local geographical source of the organism from which the sequence was derived; examples include soil, water, etc.

HTG Sequence Status

The HTG sequence status is defined by the current phase of sequence development. A sequence can be in one of the following sequence development phases:

Phase 0 (location: HTG Division): One pass read to a few pass reads of a single clone (not contigs)

Phase 1 (location: HTG Division): Unfinished, may be unordered, unoriented contigs, with gaps.

Phase 2 (location: HTG Division): Unfinished, ordered, oriented contigs, with or without gaps

Phase 3: (location: Primary Division) Finished with no gaps (with or without annotations)

There is some flexibility built into the phase definitions:

For example, although the majority of submissions represent a collection of unordered or ordered sequences derived from a single cosmid, BAC, or PAC clone, there have been cases where each individual sequence was submitted as phase 1, then updated to phase 2, then upon assembly updated to phase 3.

Modifier

Also called **Qualifier**

Modifiers provide a means of supplying extra information about a feature in addition to that provided by feature itself and location. "Source Material Modifiers" for example, would therefore be specific pieces of additional information about the source from which the sequence came.

Modifiers take the form of a slash (/) followed by the modifier name and, if applicable, an equal sign (=) and a value.

You will find a comprehensive list of modifiers (qualifiers) in section 7.3.1 of the DDBJ/EMBL/GenBank Feature Table definition document.

Path

A program path shows all the nested files that ultimately lead to where a particular program is stored in your computer

Release Date

An optional date specified upon submission for the release of the submitted records. If a release date is chosen, the sequence will be released on that date or when the accession number is published. Sequences must be released when the accession number or data is published – and this includes online publication. You must specify a release date; submissions cannot be held indefinitely pending publication.

Source

A source is a type of feature (functional group) that conveys information about the biological source(s) of the specified span of the sequence, and allows a submitter to annotate information about those biological source(s) to the sequence record.

Every sequence submission should have either a single source annotated to it that spans the entire sequence or multiple sources annotated to it, which together, span the entire sequence.

For more information on the "source" feature type, please see the alphabetic list of Feature types in section 3.2 of the DDBJ/EMBL/GenBank Feature Table definition document. The entry for "source" will provide a list of modifiers (qualifiers) that can be used with the "source" feature type.

Source Identifier

One of the values that a source modifier can take is that of a source identifier —a citation or reference number that specifically identifies the biological material from which the sequence was extracted.

For instance, the source modifier /bio_material allows the submitter to annotate onto the submitted record the specific identity of the biological material from which the nucleic acid sequenced was obtained. The value of the /bio_material modifier includes a source identifier that specifically identifies the biological material used (material ID), and can also include optional codes (institution and collection codes) that indicate where the material is currently stored:

/bio_material="[<institution-code>:[<collection-code>:]]<material_id>"

Example:

/bio_material="CGC:CB3912" The value that this /bio_material modifier takes provides the institution code CGC, indicating that the source material is housed in the Caenorhabditis Genetic Center, and then gives the specific ID CB3912 (source identifier of the material used to extract the sequence).

Source Modifier

Also called Source Qualifier

A "source modifier" provides more specific information about the source material used to obtain the sequence than the feature "source" can convey by itself.

Modifiers take the form of a slash (/) followed by the modifier name and, if applicable, an equal sign (=) and a value.

Source modifiers can take the form of text, feature labels, sequences, controlled vocabulary (predefined, authorized terms that have been preselected by a particular institution), or citation/reference numbers (source identifiers). For a list of modifier types and their definitions,

You will find a comprehensive list of modifiers (qualifiers) in section 7.3.1 of the DDBJ/EMBL/GenBank Feature Table definition document.

Span

The region (in base coordinates) where a sequence feature begins and where it ends

Specimen Voucher

A specimen voucher usually includes the collector's name and a unique number, plus the name or abbreviation of the repository (e.g. museum collection or herbarium) where the specimen is housed. Here are a few examples of specimen vouchers:

C.S. Shen 2459 (HMAS)

A.J Smith 12.iii.2002 (AMNH)
H Perrier s.n. (P)

Strain

An intraspecific group of organisms possessing distinctive traits. Cultured bacteria should include a strain designation. The strain identifier is not the same as a species epithet. A strain may be designated in any manner: by the name of an individual or locality, or by a string of numbers and/or letters.

Strain Identifier

Strain identifiers distinguish specific cultures so that the connection between a parent culture to any subsequent subculture(s) can be traced. The ability to trace this connection is important when strains differ at an

infraspecies (lower than the subspecies) level, or, in some cases, when they have been misidentified and are consequently reclassified in another species.

The strain identifier you provide will serve to distinguish your isolate from other isolates that might be obtained elsewhere. Your isolates do not need to be deposited in a culture collection in order to have a strain identifier.