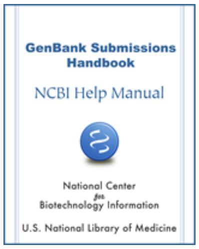
NLM Citation: The GenBank Submissions Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-. Glossary. 2011 Apr 6.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/


GenBank Submissions Handbook
NCBI Help Manual
National Center for Biotechnology Information
U.S. National Library of Medicine

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.