TnCentral Curation Guidelines

For TnCentral, we are interested in curating the following features: Protein-coding genes (e.g., transposase genes, accessory genes, passenger genes), Mobile Elements (e.g., transposons, insertion sequences, integrons), Repeat Elements (e.g., IRL, IRR), and Recombination Sites (e.g, Res and attC). These features will all be documented in the enhanced Genbank files according to the guidelines below.

General Guidelines

- In some cases, the Genbank files have additional features that we are not interested in capturing for TnCentral. Therefore, for each feature which we do want to extract, we include a field in /Note **Capture = yes**.
- Multiple items within any annotation field should be separated by "||"
- **Label** (#1 in Figure 1) refers to the text entered in the Feature text box in the SnapGene feature editing interface. **Name** (#2 in Figure 1) is a sub-field within the /note field for some feature types. For a given feature, **Name** may or may not be the same as **Label**.
- Annotation of **split features**: If a feature is interrupted by insertion of another sequence, its 5' and 3' ends should be annotated separately. For example, if the merA gene is interrupted, two features—merA 5'-end and merA 3'-end—should be created. If the feature is an ORF, "N/A" should be entered in the /product field. The 5'-end and 3'-end features should be captured (i.e., fill in Capture = yes in the /Note field) and saved in the SnapGene library. However, they should not be shown on the map. (Uncheck them in the feature list in the SnapGene file.) For the purpose of displaying in the map, a third feature should be created with the coordinates of the 5' and 3' ends shown as two disjoint regions (see **Split Feature** option below). This split feature should not be captured (i.e., do not fill in Capture = yes in the /Note field). Since it is not captured, no annotation needs to be provided other than the Label.
- Feature **Coordinates** (#3 in Figure 1) are captured automatically by SnapGene when a new feature is defined by selecting a sequence region in the SnapGene sequence or map or when a feature from the SnapGene library is detected using Detect Common Features…. If a feature is interrupted by insertion of another sequence, its coordinates can be defined as two disjoint regions using the **Split Feature** option (#4 in Figure 1, see example in Figure 2).

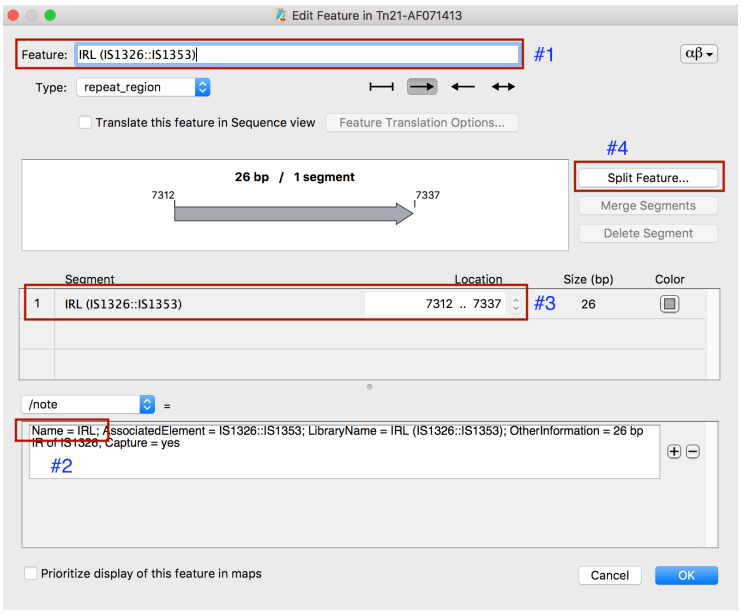

Figure 1

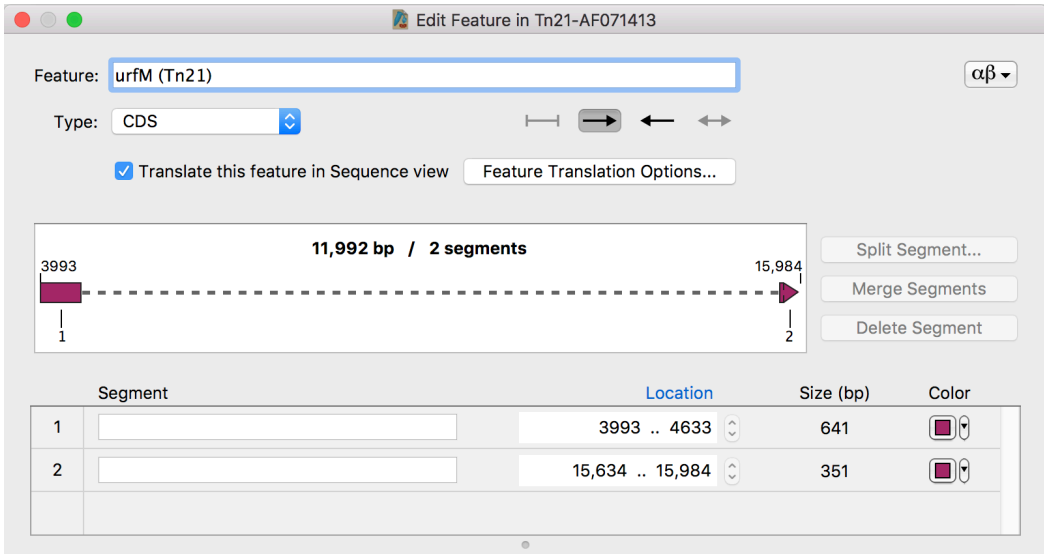

Figure 2

- Sometimes, a mobile element may contain other mobile elements inserted within it (“internal” mobile elements). Features within these internal elements should be associated with the internal element. For example, Figure 3 shows that transposon Tn1546.2 contains insertion sequence IS1216E. The tnp gene of IS1216E should be named tnp (IS1216E), not tnp (Tn1546.2).

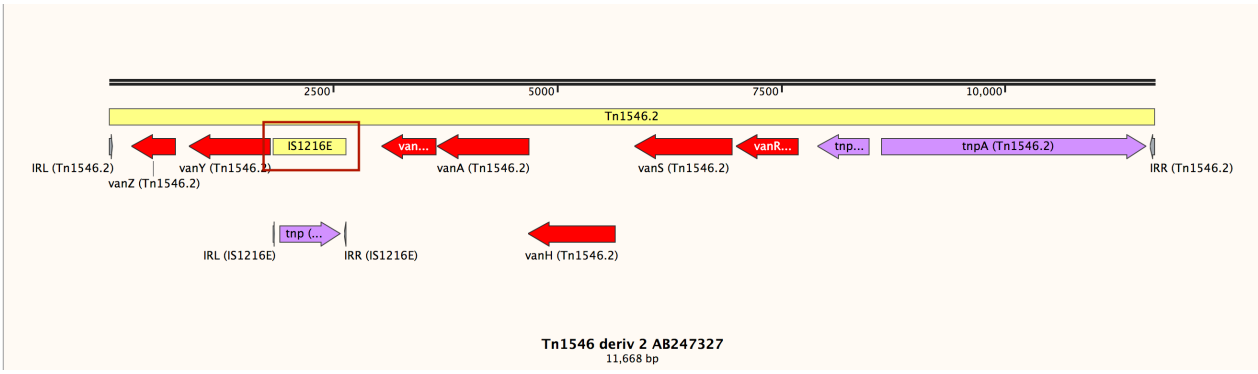

Figure 3

Mobile Element Annotation

In addition to capturing features of interest within a mobile element, the mobile element will be captured as a feature itself. This section covers annotation fields for mobile element features. Transposons should be oriented with the transposase gene oriented left-to-right. This then defines IRL and IRR.

Mobile Element Sequence Variants

There are often minor sequence variants of the same mobile element. If the sequence of the variant is >95% identical at the DNA level to the reference sequence for that element, it can be considered equivalent to an already annotated element and does not need a new name. For Insertion Sequences, the reference sequence is defined by ISfinder (https://www-is.biotoul.fr/index.php). For other elements, the reference sequence is defined by TnCentral. Variant sequences can be checked for their % identity by BLASTing against ISfinder (for Insertion Sequences) or against TnCentral (for other mobile element sequences). A separate SnapGene file should not be made for the sequence variant.

Related transposons that have slightly different complements of passenger genes are sometimes named TnXXX.1, TnXXX.2, etc. but this notation is not used consistently.

Note: Ideally, the SnapGene library will contain the reference sequences for Insertion Sequences (i.e., the sequence in the SnapGene library will exactly match the sequence in ISfinder for that Insertion Sequence. However, this is difficult to implement systematically. Therefore, some Insertion Sequence may be represented temporarily in the SnapGene library with a sequence that doesn't exactly match the reference. If a closer match the reference sequence in found during subsequent annotations, the library copy should be updated with the more closely matching sequence. However, the TnCentral Accession Number should NOT be changed. In these cases, a updated separate SnapGene file should be made for the Insertion Sequence as well.

Feature Label

The label for mobile element features will simply be the mobile element name (see below).

Feature Type (selected from SnapGene Pull-Down Menu)

mobile_element

Feature Graphic

Select the graphic with no arrowheads (Figure 4).

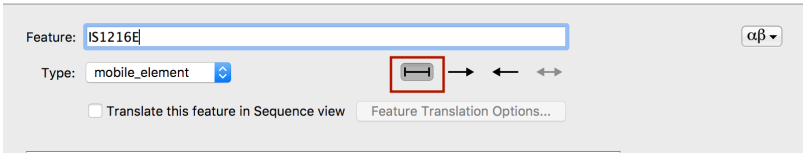

Figure 4

Feature Color

mobile element

Mobile Element Type and Name (/mobile_element_type field)

- Select type from pull-down menu (e.g., transposon, insertion sequence, integron).
- Enter the name of the mobile element (e.g., Tn21) in the text box.
- If one mobile element is interrupted by insertion of another mobile element, it will sometimes be named element1::element2. (e.g., IS1326::IS1353; this is read as "IS1326 interrupted by IS1353"). However, this convention is not always followed.
- Integrons should be named In_<parent TE name>, e.g., In_Tn7 for an integron found in Tn7 unless they are given a specific name in a paper about the Tn (e.g., In2 in Tn21, In4 in Tn1696).

Other Annotation (/note field)

- The **/note** field can contain the following sub-fields. Sub-fields should be separated with semi-colons.
    - **Accession =** Accession Number of the enhanced GenBank file for this mobile element. It is generally of the form MobileElementName-OriginalGenBankAccession, where the OriginalGenBankAccession is the GenBank Accession of the large DNA sequence (e.g., plasmid or chromosome) from which the mobile element was isolated (e.g., Tn21-AF071413)
    - **Family =** Mobile element family (e.g., Family = Tn3). For Insertion Sequences (IS), this information can often be found in ISfinder.
    - **Group =** Mobile element group (e.g., Group = Tn21). For Insertion Sequences, this can often be found in ISfinder
    - **Synonyms =** Other names for the mobile element
    - **Partial =** Enter "yes" if only a partial copy of the transposon is present
    - **Transposition =** Enter "yes", "no", "ND" (not determined) depending on whether the transposition ability of the transposon has been observed. The presence of 5bp direct

repeats just outside of the transposon sequence can be used as evidence of transposition. If the original GenBank file covers a larger region than just the transposon itself (e.g., a full plasmid sequence), then the presence of direct repeats can be checked.

- **OtherInformation =** Miscellaneous information. For Insertion Sequences, indicate the ISfinder accession number. For minor sequence variants, indicate the percent identity to the reference sequence and include the accession number of the reference sequence (e.g., OtherInformation = 99% identical to reference sequence for IS26 (ISfinder:X00011). If no % identity is noted, it is assumed to be 100% identical to reference (although 100% identity can be noted if desired.) Other miscellaneous information can also be included here.
- **Hosts:** the Hosts sub-field has several sub-sub-fields; these are separated from each other by the "|" symbol (see *Examples to Copy and Paste*). Organism and MolecularSource information should be taken from the Description Panel of the SnapGene file. Epidemiological information (e.g., Region, Country, OtherLocInfo, DateIdentified) can be taken from articles in the references.) No Host information should be recorded for Insertion Sequences. ISfinder is considered the definitive resource for the host range of Insertion Sequences.
  - **Organism =** Genus, species, and strain of the organism in which the element was identified
  - **Taxonomy =** NCBI taxon ID of the organisms in which the element was identified
  - **BacGroup =** Group (informal, not taxonomic) that the host bacterium belongs to (e.g., enterobacteria)
  - **MolecularSource =** Plasmid or chromosome on which the element was found
  - **Region =** Geographic region where the element was identified
  - **Country =** Country where the element was identified
  - **OtherLocInfo =** Other information about the location where identified
  - **DateIdentified =** Date the mobile element was identified
  - **First =** Enter "yes" if this is the first host in which the transposon was discovered
- **Capture =** Include Capture = yes for all features to be included in the database

Annotation of Internal Mobile Elements

Sometimes, a mobile element ("main" mobile element) may contain other mobile elements inserted within it ("internal" mobile elements). If an internal mobile element is already in the SnapGene library and is fully annotated, no further changes are needed. If the internal mobile element has not been previously annotated (i.e., does not have a gold star in the SnapGene library), then it should be annotated according to the guidelines for annotating mobile elements above. The host information should be the same as for the main mobile element. After annotation, regions spanned by internal mobile elements should be copied to their own SnapGene files (Figure 5). The Description Panel information of the main mobile element should be copies to the new file. Regardless of their orientation in the original transposon, in the separate files, the elements should be oriented in the conventional way with the transposase gene oriented left to right, the IRL on the left and the IRR on the right. The Flip Sequence option in the View menu of SnapGene and be used to flip the orientation of the element.

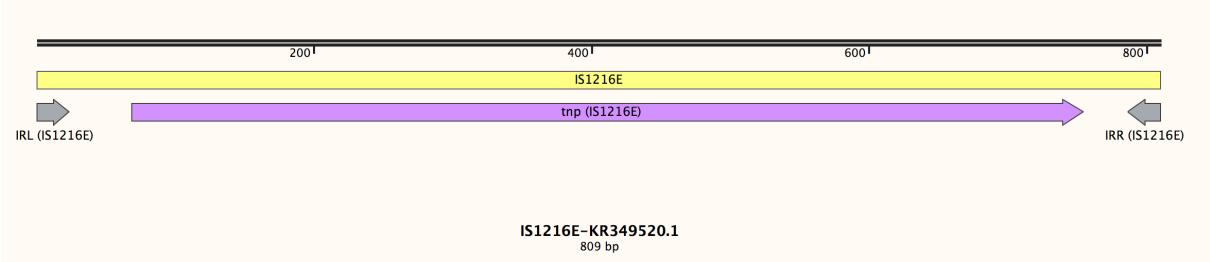

Figure 5

Annotation of Protein Coding Genes

This section covers annotation of genes that code for proteins, including transposases genes, accessory genes (e.g., resolvases), and passenger genes (e.g., antibiotic resistance, heavy metal resistance, plant pathogen, and toxin/anti-toxin genes).

Feature Label

Gene name (see below) followed by the AssociatedElement (see below) in parenthesis. For example: merA (Tn3).

Feature Color

Class = Transposase

Class = Accessory Gene (unless Subclass = CRISPR/Cas)

Class = Accessory Gene; Subclass = CRISPR/Cas

Class = Integron Integrase

Class = Passenger Gene; Subclass = Antibiotic Resistance

Class = Passenger Gene; Subclass = Heavy Metal Resistance

Class = Passenger Gene; Subclass = Plant Pathogenicity

Class = Passenger Gene; Subclass = Toxin/Antitoxin

Class = Passenger Gene; Subclass = Other

 Class = Passenger Gene; Subclass = Hypothetical

Feature Type (selected from SnapGene Pull-Down Menu)

CDS

Gene Name (/gene field)

- Use official gene symbol, when possible. By convention, gene name start with a lowercase letter (e.g., merA)
- If a gene is interrupted by insertion of another sequence, the name should be of the form geneName_disrupted (e.g., merA_disrupted).

Protein Name (/product field)

- Usually the protein name will be the same as the gene name, but starting with an uppercase letter (e.g., MerA).
- For disrupted genes, enter N/A in the /product field.

Function (/function field)

- Description of function (e.g., GO terms, UniProt keywords, Antibiotic Resistance Ontology (ARO) terms or free text). Use defined vocabularies/ontologies for annotating function whenever possible
- If terms from defined vocabularies are used, include the term identifier in parentheses after the term (e.g., antibiotic inactivation (ARO:0001004); response to mercury ion (GO:0046689)

Other Annotation (/note field)

- The **/note** field can contain the following sub-fields. Sub-fields should be separated with semi-colons. Not all sub-fields will be relevant for all genes.
    - **AssociatedElement =** the name of the mobile element that the protein-coding gene is associated with. In most cases, this will be the main mobile element being curated. However, some mobile elements contain other mobile elements inside them. In these cases, some protein-coding genes might be associated with these internal mobile elements. The Feature Label (see above) is composed of the gene name and the contents of the AssociatedElement field in parenthesis.
    - **LibraryName =** the unique identifier for this sequence in the Custom Feature library. Usually, the LibraryName will be based on the first example of the sequence found, so it will have a form that resembles a Feature Label (e.g., merA (Tn21)), but that format is not mandatory.
    - **Class** = Major classification of gene. Possible options: Transposase, Accessory Gene, Integron Integrase, Passenger Gene, Unclassified, Hypothetical
    - **Subclass** = Secondary classification of gene (see examples in section *Protein Coding Gene Classification*)

- **Target** = molecular target of the gene (e.g., type of antibiotics targeted by antibiotic resistance genes or metal targeted by heavy metal resistance genes; see examples in section *Protein Coding Gene Classification*)
- **Chemistry** = mechanism of action of transposase or resolvase genes (see examples in section *Protein Coding Gene Classification*)
- **SequenceFamily** = group or family to which gene belongs
- **OtherInformation** = miscellaneous information
- **Capture =** Include Capture = yes for all features to be included in the database

Further Information About Annotation of Some Specific Types of CDS

Accessory Genes

The fields that are typically filled in for accessory genes are: Label, /gene, /product, and /note. The sub-fields within /note that are typically filled in are: AssociatedElement, Class, Subclass, SequenceFamily, Chemistry, LibraryName, and Capture. Entries for the /gene, /product, Class, Subclass, SequenceFamily and Chemistry fields for accessory genes commonly occurring in TnCentral are shown in the file tncentral_accessory_gene_annotation.txt. Copy/paste from this file into the appropriate fields in SnapGene whenever possible.

Integron Integrases

There are three types of commonly occurring integron integrases (Class 1, Class 2, Class 3). Annotations for the /gene, /product, Class, Subclass, SequenceFamily, Chemistry, and OtherInformation fields for the three types of integrases are shown in the file tncentral_integrase_annotation.txt. Copy/paste from this file into the appropriate fields in SnapGene whenever possible.

Passenger Genes: Subclass Toxin and Antitoxin Genes

The fields that are typically filled in for toxin and antitoxin genes are: Label, /gene, /product, and /note. The sub-fields within /note that are typically filled in are: AssociatedElement, Class, Subclass, SequenceFamily, Target, LibraryName, and Capture. Entries for the /gene, /product, Subclass, SequenceFamily and Target fields for toxin and antitoxin genes commonly occurring in TnCentral are shown in the file tncentral_TA_gene_annotation.txt. Copy/paste from this file into the appropriate fields in SnapGene whenever possible.

Passenger Genes: Subclass Antibiotic Resistance (ABR) Genes

ABR genes should be annotated using the Antibiotic Resistance Ontology (ARO) as implemented in the Comprehensive Antibiotic Resistance Database (CARD; https://card.mcmaster.ca/home). BLAST the protein coding sequence of the putative ABR gene using the CARD BLAST tool. If there is a significant hit (percent identity >50% over the full length of the protein, e-value < 10^{-10}), then the ARO information for that hit should be used to annotate the ABR gene. Click on the Name or ARO tag of the hit to go to the CARD page for the gene. Information on most of the ABR genes in ARO, formatted according to TnCentral guidelines can be found in the TnCentral ABR annotation file (e.g., tncentral_abr_gene_annotation.xlsx). Whenever possible, use this file to copy/paste into the appropriate fields in SnapGene.

- Label: name of gene followed by associated element in parentheses, e.g., lsaE (TnSsu5). Do NOT include the ARO ID in this field.
- /gene: name of gene followed by ARO accession, e.g., lsaE (ARO:3003206). Copy from Gene column of TnCentral ABR annotation file.
- /product: name of gene with first letter capitalized, e.g., LsaE. Do NOT include ARO ID in this field. Copy from Product column of TnCentral ABR annotation file.
- /function: Resistance Mechanism followed by the ARO accession for the Resistance Mechanism, e.g., antibiotic target protection (ARO:0001003). Copy from Function column of TnCentral ABR annotation file.
- /note
    - AssociatedElement: as usual
    - Class = Passenger Gene
    - Subclass = Antibiotic Resistance
    - SequenceFamily: enter the AMR Gene Family term from the CARD page followed by the ARO accession for the term, e.g., ABC-F ATP-binding cassette ribosomal protection protein (ARO:3004469). Copy from SequenceFamily column of TnCentral ABR annotation file.
    - Target: enter the Drug Class terms and their ARO accessions (e.g., pleuromutilin antibiotic (ARO:3000670)||lincosamide antibiotic (ARO:0000017)||streptogramin antibiotic (ARO:0000026)|| pristinamycin IA (ARO:3000583)). Copy from Target column of TnCentral ABR annotation file. If any additional antibiotics are identified in the references for the TE, these can be included as well. The ARO accession for an antibiotic can be found by searching for the antibiotic name in CARD.
    - Chemistry: N/A
    - LibraryName: as usual
    - Other Information: indicate whether the sequence was a perfect (100%), strict, or loose match to the reference sequence for the gene. The Detection Models section of the CARD gene page lists a bitscore cutoff for a match to be considered strict. If the match is not a perfect match, indicate the bitscore. Also, list synonyms from the Synonym(s) line of the the CARD gene page. For example, strict match to reference sequence for ARO:3000596 (bitscore: 450)||Synonyms: ermCD, ermCX
    - Capture: yes

Passenger Genes: Subclass Other

This category is for CDS that have some annotation, but do not fall into any of the specific TnCentral categories.

- Color: 
- Label, /gene, /product: Possibilities are to use the name provided in the Genbank file, a name from the literature about the mobile element, the name of one of the top significant BLAST hits, or a name based on a conserved domain that the protein contains, whichever seems most

appropriate. Try to avoid really generic names like “orfX” or “urf1” (urf stands for “uncharacterized reading frame”.) Note that protein names should be succinct. Also, it is OK if the name describes the function of the protein, but it ideally shouldn't refer only to a structural domain. If the domain has some function associated with it, the information could be used in the name. For example, if the domain is generally associated with acetyltransferases, you could name the protein "putative acetyltransferase". However, in the absence of functional clues, "ABC-domain protein" is still probably a better name than "orfX." Importantly, before settling on the name, check to see if there is already an entry in the library with that same name. If there is, check if the sequences have full-length significant similarity and the same domain structure. If so, fine--use the same name. If the name is already used, but the sequences are unrelated, then you need to choose a different name.

- Class = Passenger Gene
- Subclass = Other
- SequenceFamily: enter the name(s) and accession(s) of the conserved domain(s) or famil(ies). (e.g., DUF1010 (Pfam:PF06231))
- OtherInformation = enter any additional annotation, if any

Passenger Genes: Subclass Hypothetical

This category is for ORFs that have no available annotation (i.e., all good BLAST hits are to hypothetical or uncharacterized proteins, no conserved domains or family memberships):

- Color: 
- If the ORF is < 50 amino acids and/or it overlaps with other better characterized ORFs, do not annotate it or display it in the map.
- Label, /gene, /product: Use the protein_id from the GenBank record (first choice), the locus_tag from the GenBank record (second choice) or the ID of the best scoring significant BLAST hit (third choice).
- Class = Passenger Gene
- Subclass = Hypothetical

Annotation of CDS in Mobile Element Sequence Variants

Sometimes mobile elements are encountered that are minor sequence variants of elements that are already in the TnCentral or ISfinder database (>95% sequence identity, see section on Mobile Element Sequence Variants above). In these cases, the variant is not given a new name (i.e., the name is the same as the sequence variant). This will lead to the CDS and other features of the variant having the same names as the features of reference element even though they might not be identical in sequence to the reference features. To capture these differences, note the % identity to the reference sequence at the nucleotide level in the OtherInformation field. For example, if a minor sequence variant of IS26 is found in which the DNA sequence of the tnp gene is 99% identical to the reference sequence (in ISfinder), the annotation of the tnp ORF should include OtherInformation = 99% identical to reference sequence for tnp (IS26) (ISfinder:X00011). If the protein product is severely altered by the sequence

variation (e.g., a frameshift or point mutation leads to introduction of a premature stop codon), the feature can be given a different name (e.g., tnp_p) and the nature of the change should be described in the OtherInformation field. If no % identity is noted, it is assumed to be 100% identical to the reference (although 100% identity can be noted if desired). When possible, the reference sequence of the feature should be stored in the SnapGene library. If the reference sequence is not readily available then a variant sequence can be stored temporarily.

Repeat Element Annotation

This section covers repeat elements such as IRL and IRR.

Feature Label

These features will be named by the name of the repeat element (see below) followed by the mobile element name in parentheses (e.g., IRL (Tn21) or IRR (IS1353))

Feature Type (selected from SnapGene Pull-Down Menu)

repeat_region

Feature Color

 repeat elements

Other Annotation (/note field)

- The **/note** field can contain the following sub-fields. Sub-fields should be separated with semi-colons.
    - **Name =** name of the repeat element (e.g., IRL or IRR)
    - **AssociatedElement =** mobile element with which the repeat is associated
    - **LibraryName =** the unique identifier for this sequence in the Custom Feature library. Usually, the LibraryName will be based on the first example of the sequence found, so it will have a form that resembles a Feature Label (e.g., IRL (Tn21)), but that format is not mandatory.
    - **OtherInformation** = miscellaneous information
    - **Capture =** Include Capture = yes for all features to be included in the database

Annotation of Repeat Elements in Mobile Element Sequence Variants

Annotation of repeat elements in mobile element sequence variants should be handled similarly to the annotation of CDS in sequence variants (see above). The percent identity to the reference sequence should be noted in the OtherInformation field.

Recombination Site Annotation

This section covers attC sites in Integrons and Tn3 family recombination sites (Res). Res sites also contain subsites with agreed names. Integron attC sites are disrupted by insertion of a gene cassette.

Therefore, they should be annotated according to the procedures for split features (annotate the 5'-end and 3'-end as well as a split feature for display on the map).

Feature Label

The label will be the name of the site (see below) followed by the mobile element name in parentheses (e.g., res_site_III (Tn21))

Feature Type (selected from SnapGene Pull-Down Menu)

misc_feature

Feature Color

 recombination site

Other Annotation (/note field)

- The **/note** field can contain the following sub-fields. Sub-fields should be separated with semi-colons.
    - **Name =** name of the recombination site (e.g., res_site_I). Integron attC sites should be named "attC-<name of downstream gene>" (e.g., attC-aadA)
    - **AssociatedElement =** mobile element with which the recombination sites is associated
    - **LibraryName =** the unique identifier for this sequence in the Custom Feature library. Usually, the LibraryName will be based on the first example of the sequence found, so it will have a form that resembles a Feature Label (e.g., res_site_I (Tn21)), but that format is not mandatory.
    - **Capture =** Include Capture = yes for all features to be included in the database

Annotation of Recombination Sites in Mobile Element Sequence Variants

Annotation of recombination sites in mobile element sequence variants should be handled similarly to the annotation of CDS in sequence variants (see above). The percent identity to the reference sequence should be noted in the OtherInformation field.

Field Structure for /note for Different Feature Types

Below is the field structure for /note for different feature types. It is not required to fill in all fields in all cases.

Mobile Elements

Accession = ; Family = ; Group =; Synonyms =; Partial = ; Transposition = ; OtherInformation = ; Hosts: Organism = |Taxonomy = |BacGroup = |MolecularSource = |Region = |Country = |OtherLocInfo = | |First = ; Capture = yes

CDS

AssociatedElement = ; Class = ; Subclass = ; SequenceFamily = ; Target = ; Chemistry = ; LibraryName = ; OtherInformation = ; Capture = yes

Repeat Regions and Recombination Sites

Name = ; AssociatedElement = ; LibraryName = ; OtherInformation = ; Capture = yes