

## SOP for Annotation of Insertion Sequences

<b>ERIC-BRC UW Team- Standard Operating Procedure</b>			
<b>UWSOP#A005</b>			
<b>Title: IS Elements</b>			
<b>Purpose: Annotation of Insertion Sequences</b>			
<b>Expected Result of Following Procedure: Replace with Result</b>			
<b>Version#</b>	<b>Date</b>	<b>Author</b>	<b>Reviewer</b>
<b>1</b>	<b>12/13/2006</b>	<b>Val Burland</b>	<b>Not Reviewed Yet</b>

- 1. Search** Genome sequences are searched against a database of IS elements by Repeat Masker (Washington U website; Bioinformatics. 2000 Nov. 16(11):1040-1), enclosed in a wrapper script and running on the ASAP server.
- 2. Processing results** The text output file is modified to tab-delimited text for import into Excel. Alignment length and IS length are computed, and then used to calculate the proportion of the IS included in the alignment.
- 3. Criteria for annotation** IS and fragments are annotated if the alignment length is >100 bp or >10% of the element. Data in the table is easily sorted to prepare annotations.
- 4. Inspection of alignments** The output data includes % divergence and indels; adjacent or nearby hits may be different parts of the same element. Pairwise alignments (e.g. using Lasergene Megalign) are used to assess the quality of the hit and capture useful information to be added as a “note” qualifier. When pre-existing IS annotations do not match RepeatMasker output, alignment inspections are needed to resolve discrepancies.
- 5. Preparing annotations** IS elements are currently annotated as “repeat\_region” (feature type) with “insertion sequence” qualifiers. The qualifier data is the IS name and evidence is NSS (nucleotide sequence similarity) and ISN (IS name) which creates a link to the relevant page in the IS database ISFinder (home page <http://www-is.biotoul.fr/is.html>).

Incomplete elements are annotated in the same way, but a “note” qualifier is added using the term “fragment”, and describing the nature of the fragment (examples below). The qualifier “partial” should not be used.

Multiple elements of one type are distinguished by the qualifier “name” consisting of the IS name followed by a dot and number suffix, .1, .2, .3, etc. Single elements are also named, followed by .1. (see examples). Full chromosomal elements are numbered first, then incomplete chromosomal elements, then plasmids.

### 6. Examples:

Full length IS:

```
repeat_region
insertion sequence      IS285 evidence (NSS ISN  ISname)
name IS285.13          evidence (USA ANEC  annotator name; email address)
```

Incomplete IS

```
repeat_region
insertion sequence      IS1441
name IS1441.23
note 5' fragment, internal 14 bp deletion and truncated by IS1F
```

or note internal fragment  
or note fragment, internal deletions of 25 and 49 bp.

(for multipart IS features) note disrupted by insertion of IS4  
Evidence for the note is NSS ISN ISname.

**Footnote:** GenBank proposes to change the form of mobile DNA feature annotations in December 2006. The new format will be:

/mobile element= insertion sequence:IS289

