Proper Use of /locus_tag in Genome Submissions

At the International Nucleotide Sequence Database Collaborators meeting, it was agreed that we would require genome projects to be registered with the database. Each genome project would be assigned an ID in order to allow us to associate multiple sequences of a single genome project with each other. This Genome Project ID will appear in a new line type below ACCESSION and VERSION in the flat file. Registration of Genome Projects can be done at DDBJ, EBI or NCBI. A submitter can also register for a locus_tag prefix at the same time that they register their genome project.

Locus_tags are identifiers that are systematically applied to every gene in a genome. These tags have become surrogate gene names by the biological community. If two submitters of two different genomes use the same systematic names to describe two very different genes in two very different genomes, it can be very confusing. In order to prevent this from happening INSD has created a registry of locus_tag prefixes. Submitters of eukaryotic and prokaryotic genomes should register their prefix prior to submitting their genome. All components of a project (such as multiple chromosomes or plasmids, etc) should use the same locus_tag prefix.

The locus_tag prefix can contain only alpha-numeric characters and it must be at least 3 characters long. It should start with a letter, but numerals can be in the 2nd position or later in the string. (ex. A1C). There should be no symbols, such as -_* in the prefix. The locus_tag prefix is to be separated from the tag value by an underscore '_', eg A1C_00001.

Locus_tags should be assigned to all protein coding and non-coding genes such as structural RNAs. /locus_tag should appear on gene, mRNA, CDS, 5'UTR, 3'UTR, intron, exon, tRNA, rRNA, misc_RNA, etc within a genome project submission. Repeat_regions do not have locus_tag qualifiers. The same locus_tag should be used for all components of a single gene. For example, all of the exons, CDS, mRNA and gene features for a particular gene would have the same locus_tag. There should only be one locus_tag associated with one /gene, i.e. if a /locus_tag is associated with a /gene symbol in any feature, that gene symbols (and only that /gene symbol) must also be present on every other feature that contains that locus_tag.

Locus_tags are systematically added to genes within a genome. They are generally in sequential order on the genome. If a genome center were to update a genome and provide additional annotation, the new genes could either [1] be assigned the next sequential available locus_tag or [2] the submitter can leave gaps when initially assigning locus_tags and fill in new annotation with tag values that are between the gaps.

Use:

Incremental locus_tags

Original submission	Revised submission

ABC_0022	ABC_0022
	ABC_4568 (new gene)
ABC_0023	ABC_0023

OR

Gaps in original locus_tags

Original submission	Revised submission
ABC_0020	ABC_0020
	ABC_0021 (new gene)
ABC_0030	ABC_0030

BUT NOT

Decimal integers

Original submission	Revised submission
ABC_0020	ABC_0020
	ABC_0020.1 (new gene)
ABC_0030	ABC_0030

It is preferable to use the same numbering convention for all locus_tags within a project no matter whether the gene is a protein coding gene or structural RNA or from one chromosome or another.

However, submitters wishing to encode information about chromosome number, or RNA type in the locus_tag value, may add this information to the /locus_tag after the prefix and underscore:

ABC_I00001 for gene 1, chromosome I
ABC_II00001 for gene 1, chromosome II
ABC_r1112 for ribosomal RNA genes
ABC_t1113 for tRNA genes

A submitter can register for a locus_tag prefix and project ID at NCBI , EBI or DDBJ. It is preferable that you register for your project ID and locus_tag prefix at the site where you intend to submit your genome; do not register at all three sites. When a locus_tag prefix request is submitted to the database, there is a check to see whether that prefix has already been registered to another project. If the prefix is available, the submitter is informed that this locus_tag is registered for their project. If it is not available, the interface will report that this locus_tag has already been taken. The submitter can then choose to check for another prefix or to have the database suggest an unregistered prefix for the project.