CLdb (CRISPR Loci Database) tutorial

Example run

Setup

`CLdb setup requires the following files:`

- Loci table (tab-delimited); columns need:
  - Taxon_ID
  - Taxon_Name
  - Subtype
  - Locus_Start
  - Locus_End
  - Operon_Start
  - Operon_End
  - CRISPR_Array_Start
  - CRISPR_Array_End
  - Status
  - Genbank
  - Array_File
  - Author
  - File_Creation_Date
- Array table files (tab-delimited; copy and paste from CRISPRFinder); columns needed:
  - Start position
  - Direct repeat sequence
  - Spacer sequence
  - End position
- Genbank files for each organism of interest
  - merged
  - FIG-PEG IDs for CDS features in db_xref tags (e.g. "fig|6666666.40253.peg.2362")

EXAMPLE RUN

Directory setup

```
The directory name for this example: './CLdb/'
The example loci table: 'loci.txt'

$ mkdir CLdb
$ cd CLdb
$ mkdir genbank
    # place/symlink genbank files in this directory
$ mkdir array
    # place/symlink array files in this directory
```

making the tables in the database

```
$ CLdb_makeDB.pl -r
```

loading the loci table

```
$ CLdb_loadLoci.pl -d CRISPR.sqlite < loci.txt
```

adding number of scaffolds to the loci table

```
$ CLdb_addScaffolds.pl -d CRISPR.sqlite
```

loading arrays and direct repeats to their respective tables

```
$ CLdb_loadArrays.pl -d CRISPR.sqlite
```

grouping spacers and direct repeats (groups with same sequence)

```
$ CLdb_groupArrayElements.pl -d CRISPR.sqlite -s -r
```

getting genes in CRISPR locus region (defined in Loci table)

```
$ CLdb_getGenesInLoci.pl -d CRISPR.sqlite > gene_table.txt
    # <optional> manually currate the 'gene_alias' column values
```

loading genes into the Genes table

```
$ CLdb_loadGenes.pl -d CRISPR.sqlite < gene_table.txt
```