**Characterization of RVDB sequences**

1. **Overview**

The “RVDB\_characterization.py” script was used to partition the sequences into five Level 1 categories: exogenous viral, endogenous nonretroviral, endogenous retroviral, LTR-retrotransposon, and unassigned viral gene /fragments. This partitioning was done using some of the SEM-R positive keywords, and organizing them by categories. Sequences possessing headers with specific positive keywords from SEM-R screen were placed into the corresponding categories. For instance, the keywords “retrotranspos”,”retro transpos”,”retroelem”,”blastopia “,” copia “,” delta element”,” gipsy “,” gypsy element”,” gypsy like “,” gypsy type “, “insertion element”, ” mdg1 “, ” mdg3 “, ”micropia”, “ sire “,” ty element”, and “ ty insertion” were used to classify sequences as belonging to the LTR-retrotransposon category. There is also a regular expression for finding strings of the form “ ty” + either “1” or “3”, with / without a space, and there are also a handful of rules for pulling in the less common LTR-retrotransposons: the string “transpos” + either “ bel “, “ pao “, “ roo “, or “morgane”. There are similar combinations of keywords, regular expressions, and rules for the other four groups.

1. **Running RVDB\_characterization.py**

The RVDB\_characterization.py script can be run by a single line in the command shell, containing python command, the script name, and then 5 parameters: the home directory (one level below the update folder), the date tag for the update, the current version of the update, the name of the fasta file to be characterized (e.g. “U-RVDBv13.0.fasta”), and a filename containing a filterset, an accession list for a subset of sequences to be characterized. The last two parameters can be selected so that the script can be run not just on the base unclustered form of RVDB, but also the clustered form of RVDB, or any special-purpose sub-version created by the user. Please note, if all sequences from the supplied fasta file are to be characterized, there is no filterset and the final parameter can be a random letter (e.g. “NA”). Below is an example of running the script from windows.cmd:

>python E:/UPDATE\_SCRIPTS\_LOGS/RVDB\_characterization.py E: apr.2018 13.0 U-RVDBv13.0.fasta NA

The counts for each group are recorded in a log file, $fastafilename”\_char\_output\_log.txt”, which is in the current update folder. For example:

E:/RVDBv13.0/U-RVDBv13.0\_char\_output\_log.txt

In addition, the script generates output files of headers for each of the categories. These are also written to the current update folder and are named $fastafilename”.”$group”.headers.txt”, for example:

E:/RVDBv13.0/RVDBv13.0.fasta.EX.headers.txt

E:/RVDBv13.0/RVDBv13.0.fasta.ENRV.headers.txt

E:/RVDBv13.0/RVDBv13.0.fasta.ERV.headers.txt

E:/RVDBv13.0/RVDBv13.0.fasta.LTR-RETO.headers.txt

E:/RVDBv13.0/RVDBv13.0.fasta.UNASSIGNED.headers.txt

1. **Manual Review**

In our characterization efforts, we did find it necessary to perform some final manual review. In particular, we found that some sequences that had been labelled by the RVDB\_characterization.py script as endogenous nonretroviral, were in fact endogenous retroviral according to their name. Also, some of the sequences that had been labelled by the script as “viral gene/fragment” were in fact exogenous viral or LTR-retrotransposon.