This folder contains the scripts for creating a control database of random insertions to investigate the enrichment of retrotransposable elements (RTEs) in the euchromatin domains of 127 cell types analysed by the Roadmap project (Roadmap Epigenomics Consortium, 2015). Accessible chromatin regions identified by ChIP-seq analysis for histone marker H3K4me3, associated with euchromatin domains for each of the 127 cell types were obtained from the roadmap project repository (https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/; last modified: October 2013) and concatenated in shell.

The Roadmap file of euchromatin regions is available in OneDrive via the following link:

<https://liveplymouthac-my.sharepoint.com/:u:/g/personal/randa_ali_plymouth_ac_uk/EQSSzURYxA9GpDYA_mASn9gBgygw2hepkTA-FjItXkml3Q?e=7wVviq>

Download the compressed file, and add it to the following directory:

Path\_to\_MyAnalysis/Enrichment\_in\_Euchromatin\_regions/Raw\_data

A million random insertions are produced within the script using the default settings of the random tool of BEDtools suite version 2.25.0 (Quinlan, 2014). BEDtool random default setting creates 1 million random insertions that are 100 base pairs (bp) in length. Next, a loop function is set to subsample a set of insertions based on the size of non-reference RTEs and the frequency of random insertions interrupting a gene or enhancer region is calculated. The loop function is reiterated a 1,000 x before it terminates and return the collated results.

Euchromatin\_Enrichment\_1kRandomDataset.R: R code for creating a control dataset and analysing for 1,000 reiterations the distribution of random insertions in the euchromatin domains of 127 cell types. There are three similar scripts with subsamples proportional to the size of each RTE type.

The scripts are executable from the command line from the “Path\_To\_MyAnalysis/MyAnalysis/Shell Scripts” directory:

Sample file (5 iterations):

$ nohup Rscript Demo\_EuchromatinEnrichment\_RandomDataset.R “Path\_to\_MyAnalysis/MyAnalysis” > Euchromatin.out 2>&1 &

RTE files (1,000 iterations):

$ nohup Rscript L1\_Euchromatin\_Enrichment\_1kRandomDataset.R “Path\_to\_MyAnalysis/MyAnalysis” > L1\_Euchromatin.out 2>&1 &

$ nohup Rscript Alu\_Euchromatin\_Enrichment\_1kRandomDataset.R “Path\_to\_MyAnalysis/MyAnalysis” > Alu\_Euchromatin.out 2>&1 &

$ nohup Rscript SVA\_Euchromatin\_Enrichment\_1kRandomDataset.R “Path\_to\_MyAnalysis/MyAnalysis” > SVA\_Euchromatin.out 2>&1 &

**Output files:** Demo\_RandomH3K4Euchromatin.txt; RandomL1Hs\_1k\_H3K4Euchromatin.txt; RandomAlu\_1k\_H3K4Euchromatin.txt; RandomSVA\_1k\_H3K4Euchromatin.txt

The output of the script is a tab delimited text file consisting of 1,001 columns. The first column listing the epigenomic identifier (EID) of each cell type, and the subsequent 1,000 columns list the counts of random insertions overlapping with the euchromatin domains of each cell type for each of the 1,000 iterations.

**Statistical analysis and displaying the results:**

EnrichmentInEuchromatinRegions\_statAndPlot.R: R code for calculating the empirical P-value for the enrichment of RTEs in the euchromatin domain of 127 Roadmap cells and displaying the count of cell types enriched with one or more RTEs by tissue group. The script is executable in RStudio from the “Path\_To\_MyAnalysis/MyAnalysis/RStudio\_Scripts” directory.

**Input file:** RandomL1Hs\_1k\_H3K4Euchromatin.txt, RandomAluY\_1k\_H3K4Euchromatin.txt, RandomSVA\_1k\_H3K4Euchromatin.txt; TotalRTEsInEuchromatinRegions.txt

The input files are the output files from Euchromatin\_Enrichment code plus a tab delimited text file in table format consisting of counts of non-reference RTEs in euchromatin regions.

TotalRTEsInEuchromatinRegions.txt: A tab delimited text file containing the count of non-reference L1Hs, AluY, and SVA\_E/F elements in the euchromatin domain of each of the 127 cell types. The counts were identified by overlapping the position of RTEs with euchromatin regions using the intersect tool of BEDtools suite version 2.25.0 (Quinlan, 2014) with the following parameters:

-a: RTE features in bed file format.

-b: Roadmap peaks position in bed file format.

-wa: Write the original RTE features in -a for each overlap.

-wb: Write the original Roadmap feature in -b for each overlap.

Command:

$ bedtools intersec -a RTE.bed\* -b RoadmapAll\_H3K4me3.broadPeak\_sorted\_27062019.bed –wa –wb

Next, total number of overlap between each EID and RTE were counted in UNIX.

**Output files:** RTEEnrichmentInEuchromatin\_PvalByEID.txt; RTEEnrichmentInEuchromatin\_CountsByAnatomyGroup.txt.

RTEEnrichmentInEuchromatin\_PvalByEID.txt: A tab delimited text file in table format containing the P-value of enrichment for EIDs enriched with RTEs. The file consists of 7 columns: the first column is the EID identifications and the subsequent columns contain the RTE type and P-value of enrichment.

RTEEnrichmentInEuchromatin\_CountsByAnatomyGroup.txt: A tab delimited text file in table format containing the count of EIDs in each anatomical region for each of the RTE types. The file consists of 4 columns: the first column is the Anatomical region and the subsequent columns contain the counts of EIDs for each RTE type.

\*RTE.bed are 3 non-reference L1s, Alus, and SVA files prefixed with RTEdb\_\* available in “MyAnalysis/RTE\_files” folder.