This folder contains the scripts for creating a control database of random insertions to investigate the enrichment of retrotransposable elements (RTEs) in functional genomic regions including gene and enhancer regions. 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.

GenicEnrichment\_1kRandomDataset.R: R code for creating a control dataset and analysing the distribution of random insertions in genic vs. intergenic regions for 1,000 reiterations. 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\_genicEnrichment\_RandomDataset.R “Path\_to\_MyAnalysis/MyAnalysis” > GenicEnrichment.out 2>&1 &

RTE files (1,000 iterations):

$ nohup Rscript L1GenicEnrichment\_1kRandomDataset.R “Path\_to\_MyAnalysis/MyAnalysis” > L1\_GenicEnrichment.out 2>&1 &

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

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

**Output files**: Demo\_RandomInsertionsIngeneregions.txt; L1\_1kRandomInsertionsIngeneregions.txt; AluY\_1kRandomInsertionsIngeneregions.txt; SVAEF\_1kRandomInsertionsIngeneregions.txt

The output of the script is a tab delimited text file consisting of 3 columns, each representing the frequency distribution of random insertions in each genomic region: intergenic, intronic, or exonic.

EnhancerEnrichment\_1kRandomDataset.R: R code for creating a control dataset and analysing the distribution of random insertions in enhancer vs. non-enhancer regions for 1,000 reiterations. 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\_EnhancerEnrichment\_RandomDataset.R “Path\_to\_MyAnalysis/MyAnalysis” > EnhancerEnrichment.out 2>&1 &

RTE files (1,000 iterations):

$ nohup Rscript L1EnhancerEnrichment\_1kRandomDataset.R “Path\_to\_MyAnalysis/MyAnalysis” > L1\_EnhancerEnrichment.out 2>&1 &

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

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

**Output files**: Demo\_RandomInsertions\_Enhancerregions.txt; L1\_1kRandomInsertions\_Enhancerregions.txt; AluY\_1kRandomInsertions\_Enhancerregions.txt; SVAEF\_1kRandomInsertions\_Enhancerregions.txt

The output of the script is a tab delimited text file consisting of 1 column representing the frequency of random insertions in enhancer regions.

**Statistical analysis:**

The following R scrips are executable in RStudio from the “Path\_To\_MyAnalysis/MyAnalysis/RStudio\_Scripts” directory:

ZscoreCalculationForGenicRegions.R: R code for investigating the overrepresentation of reference and non-reference RTEs in gene regions compared with the distributions of the control dataset using Z-statistics.

**Input files**: GeneRegionsFrequencies.txt; L1\_1kRandomInsertionsIngeneregions.txt; AluY\_1kRandomInsertionsIngeneregions.txt; SVAEF\_1kRandomInsertionsIngeneregions.txt

The input files are the output files of the GenicEnrichment R codes plus a tab delimited text file in table format consisting of the frequency distribution of reference and non-reference RTEs in each genomic region (previously mentioned in FunctionalRegionsAnalysis\_README)

ZscoreCalculationForEnhancerRegions.R: R code for investigating the overrepresentation of reference and non-reference RTEs in enhancer regions compared with the distributions of the control dataset using Z-statistics.

**Input files**: EnhancerRegionsFrequencies.txt; L1\_1kRandomInsertions\_Enhancerregions.txt; AluY\_1kRandomInsertions\_Enhancerregions.txt; SVAEF\_1kRandomInsertions\_Enhancerregions.txt

The input files are the output files of the EnhancerEnrichment R codes plus a tab delimited text file in table format consisting of the frequency distribution of reference and non-reference RTEs in enhancer regions (previously mentioned in FunctionalRegionsAnalysis\_README)

**Output files:** Zstatistics\_RTEsInGeneRegionsVsRandom.txt; Zstatistics\_RTEsInEnhancerRegionsVsRandom.txt

The output files for both R codes is a tab delimited table in text file format consisting of 5 columns in the following order:

Column 1: Region. Either intergenic, intronic, or exonic for gene regions analysis, or enhancer for enhancer regions analysis.

Column 2: RTE Category. Either reference or non-reference.

Column 3: RTE type. Either L1, Alu, or SVA.

Column 4: Z-score.

Column 5: P-value (two-sided).