This folder contains the script for the distribution of retrotransposable elements (RTEs) in functional regions including gene and enhancer regions. These scripts process tab delimited text files in table format and creates bar charts that presents the frequency distribution of RTEs to facilitate visualising the data.

BarplotForGeneDis.R: R code for plotting the distribution of reference and non-reference RTEs in genic vs. intergenic regions. The script is executable in RStudio from the “Path\_To\_MyAnalysis/MyAnalysis/RStudio\_Scripts” directory:

**Input file:** GeneRegionsFrequencies.txt

The input file is a tab delimited text file in table format consisting of 4 columns in the following order:

Column 1: Genomic region. Either intergenic, intronic, or exonic.

Column 2: Frequency distribution of reference RTEs in each genomic region.

Column 3: Frequency distribution of non-reference RTEs in each genomic region.

Column 4: RTE type. Either L1, Alu, or SVA.

The frequency distribution of each RTE type in each genomic region was determined by overlapping the position of RTEs with RefSeq genes. The annotated NCBI RefSeq genes were downloaded from the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables; Karolchik, 2004) (last update 11.09.2017); under the group ‘genes and gene predictions’ and table ‘UCSC RefSeq (refGene)’. The RefSeq genes files are available in “MyAnalysis/Functional\_analysis/RefSeq” folder.

Bedtools analysis was conducted in two steps:

Step 1: The position of RTEs were overlapped with RefSeq genes using transcript start and transcript end positions 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: Gene features in bed file format.

-c: Report the number of hits in –b for each RTE entry in –a file. Reports 0 for RTE entries that have no overlap with gene regions. These insertions were considered intergenic.

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

Command:

$ bedtools intersec -a RTE.bed\* -b RefSeq\_headless\_nochr\_chr1toY\_sorted\_21092017.bed -c -wa

Step 2: RTE insertions overlapping with gene regions were then overlapped with RefSeq gene intervals file consisting of exon and intron start and end positions using the intersect tool of BEDtools suite version 2.25.0 (Quinlan, 2014) with the following parameters:

-a: RTE features with gene overlaps from step 1 in bed file format.

-b: Gene segment features (i.e. Introns, exons) in bed file format.

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

-wb: Write the original gene segment features in -b for each overlap

Command:

$ bedtools intersect -a Genic\_RTEs.bed -b Refseq\_geneIntervalsWithGeneNames\_02102017.bed -wa –wb

Frequency of insertions per genomic region was then calculated as:

BarplotForEnhancerDis.R: R code for plotting the distribushion of reference and non-reference RTEs in known enhancers vs. enhancer-free regions. The script is executable in RStudio from the “Path\_To\_MyAnalysis/MyAnalysis/RStudio\_Scripts” directory:

Input file: EnhancerRegionsFrequencies.txt

The input file is a tab delimited text file in in table format consisting of 4 columns in the following order:

Column 1: Genomic region. Either enhancer or enhancer free.

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

Column 3: Frequency distribution of each RTE type in each genomic region.

Column 4: RTE type. Either L1, Alu, or SVA.

The frequency distribution of each RTE type in enhancer regions was determined by overlapping the position of RTEs with the position of enhancers from the GeneHancer database. The GeneHancer database (Fishilevich et al., 2017; last updated 2-09-2018) was obtained via the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables; Karolchik, 2004), selecting the ‘GH Reg Elems (geneHancerRegElements)’ table as part of the ‘GeneHancer’ track under the ‘Regulation’ group. The GeneHancer file is available in “MyAnalysis/Functional\_analysis/GeneHancer” folder.

The position of RTEs were overlapped with enhancer 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: GeneHancer features in bed file format.

-wao: Reports the original RTE entry and the enhancer entry it overlaps with plus the number of base pairs of overlap between the two features. However, it also reports RTE entries without any enhancer overlaps and 0 for the number of base pairs of overlap.

Command:

$ bedtools intersec -a RTE.bed\* -b AllChromosomes\_GeneHancerRegulators\_26062019.bed –wao

Frequency of insertions enhancer regions was then calculated as:

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