This folder contains the script for the GC content analysis of the hg19 genome plus reference and non-reference RTEs in 20 kilobase (kb) windows.

The 20 kb windows for the genome file were created using the makewindows tool of BEDtools suite version 2.25.0 (Quinlan, 2014) with the following parameters:

-g: A tab delimited .genome file containing chromosome sizes for GRCh37/hg19 human genome build. Chromosome sizes were obtained from the Genome Reference Consortium (GRC) website (<https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh37>)

-w: Window size in base pairs (bp).

Command:

$bedtools makewindows -g chr\_length.genome\* -w 20000

BEDtools slop (Quinlan, 2014) restricts the resizing of each feature to the size of the chromosome. The 20kb windows for reference and non-reference insertions were created using the slop tool of BEDtools suite version 2.25.0 (Quinlan, 2014) with the following parameters:

-i: A tab delimited input file in bed format

-g: A tab delimited .genome file containing chromosome sizes for hg19 human genome build.

-b: Increases the bed file entries by the same number of bps in each direction.

Command:

$bedtools slop -i RTE\_Input.bed\* -g chr\_length.genome -b 10000

**GC content analysis:**

Note: The genome .fa file is available in OneDrive via the following link:

<https://liveplymouthac-my.sharepoint.com/:u:/g/personal/randa_ali_plymouth_ac_uk/EfsnaCycpyhIjTgF5k314u8BOuHDWlbo-ISblv78RHPXTg?e=7EtfFZ>

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

Path\_to\_MyAnalysis/GC\_content\_analysis/Raw\_data

GCcontent\_Analysis.R: R script for the GC content analysis executable from the command line from the “Path\_To\_MyAnalysis/MyAnalysis/Shell Scripts” directory:

Sample file:

Rscript Demo\_GCcontentAnalysis.R “Path\_to\_MyAnalysis/MyAnalysis”

Genome and RTE files:

Genome : $Rscript Genome\_GCcontent\_Analysis.R “Path\_to\_MyAnalysis/MyAnalysis”

Reference L1s: $Rscript RefL1\_GCcontent\_Analysis.R “Path\_to\_MyAnalysis/MyAnalysis”

Reference Alus: $Rscript RefAlu\_GCcontent\_Analysis.R “Path\_to\_MyAnalysis/MyAnalysis”

Reference SVAs: $Rscript RefSVA\_GCcontent\_Analysis.R “Path\_to\_MyAnalysis/MyAnalysis”

Non-reference L1s: Rscript NonRefL1\_GCcontent\_Analysis.R “Path\_to\_MyAnalysis/MyAnalysis”

Non-reference Alus: Rscript NonRefAlu\_GCcontent\_Analysis.R “Path\_to\_MyAnalysis/MyAnalysis”

Non-reference SVAs: Rscript NonRefSVA\_GCcontent\_Analysis.R “Path\_to\_MyAnalysis/MyAnalysis”

This code calculates the GC content of each feature and outputs a text file with the same number of rows as the input file. This section requires the human genome assembly in FASTA format. The hg19/GRCh37 human genome assembly was downloaded from the UCSC genome browser ftp website (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/) as a compressed folder (chromFa.tar.gz). Individual fasta files within the decompressed folder were concatenate in UNIX.

**Input files**:

Sample: SampleFile.bed

Genome: genomeWindow\_20kb.bed

Reference L1: L1PA2\_L1PA5\_refL1s\_20kbwindow.bed

Reference Alu: AluJ\_refAlus\_20kbwindow.bed

Reference SVA: SVAABC\_refSVAs\_20kbwindow.bed

Non-reference L1: RTEdb\_L1Hs\_all\_200bpmerged\_20kbWindow.bed

Non-reference Alu: RTEdb\_Alu\_all\_200bpmerged\_20kbWindow.bed

Non-reference SVA: RTEdb\_SVA\_all\_200bpmerged\_20kbWindow.bed

**Input file format**:

The input file for the R code is a tab delimited bed file where the first three fields in each feature line are required:

1. Chromosome name which can be given with or without the 'chr' prefix.
2. Start position of the feature in standard chromosomal coordinates.
3. End position of the feature in standard chromosomal coordinates.

Eight additional optional columns are included in the non-reference L1s, Alus, and SVAs files:

Strand, RTE type, RTE subfamily, Allele frequency, Total number of individuals carrying the insertion, Study reference, Study PubMed ID.

Three additional optional columns are included in the reference L1s, Alus, and SVAs files:

RTE subfamily, REF, strand.

**Output files:**

Sample: Demo\_GCcontent.txt

Genome: GCcontent\_wholeGenome\_20kb\_09022018.txt

Reference RTEs: RefL1PA2PA5\_GCresults20kb\_29052019.txt; RefAluJ\_GCresults20kb\_29052019.txt; RefSVAAC\_GCresults20kb\_29052019.txt

Non-reference RTEs: RTEdb\_L1Hs\_GCresults20kb\_26072019.txt; RTEdb\_AluY\_GCresults20kb\_26072019.txt; RTEdb\_SVAEF\_GCresults20kb\_26072019.txt

These are tab delimited text files consisting of two columns. The first column is the row number of the feature from the input file and the second column is the GC-content of the feature calculated as the percentage of GC nucleotides out of total nucleotides.

**GC content figure plotting and statistical analysis:**

GCcontent\_figure\_stat.R: R script for plotting the GC distribution in a linear graph and calculating statistical significance. The script is executable in RStudio from the “Path\_To\_MyAnalysis/MyAnalysis/RStudio\_Scripts” directory:

L1s: L1\_GCcontent\_figure\_stat.R

Alus: Alu\_GCcontent\_figure\_stat.R

SVAs: SVA\_GCcontent\_figure\_stat.R

Each code requires 3 input files of the genome, reference and non-reference RTEs.

The code is divided into two sections:

**Section 1** plots a linear graph overlaying the GC-content distribution of the genome with the GC-content distributions of reference and non-reference RTEs in the same figure to facilitate comparison.

**Section 2** compares the GC-content distributions using the nonparametric Kolmogorov–Smirnov (K-S) test.

**Input files**:

The input files for this code is the output files from GC content analysis codes:

Genome: GCcontent\_wholeGenome\_20kb\_09022018.txt

Reference L1s: RefL1PA2PA5\_GCresults20kb\_29052019.txt

Reference Alus: RefAluJ\_GCresults20kb\_29052019.txt

Reference SVAs: RefSVAAC\_GCresults20kb\_29052019.txt

Non-reference L1s: RTEdb\_L1Hs\_GCresults20kb\_26072019.txt

Non-reference Alus: RTEdb\_AluY\_GCresults20kb\_26072019.txt

Non-reference SVAs: RTEdb\_SVAEF\_GCresults20kb\_26072019.txt