Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/13470
DOI URL: http://dx.doi.org/10.21474/IJAR01/13470

REVIEW ARTICLE

SIMPLE R TOOLS FOR GENETIC MARKERS RESEARCH

Victor Sitnic, Valentina Stratan, Valeri Tutuianu, Cristina Popa and Veronica Balan

Manuscript Info

Manuscript History
Received: 28 July 2021
Final Accepted: 31 August 2021
Published: September 2021

Key words:-
R. Language, Fasta Sequences, Genomic Data, Molecular Markers

Introduction

R is a free licensed programming language which presents a big interest as a tool for bioinformatics data analysis. It is essential in research activities related to the analysis of molecular-biological data and the identification of molecular markers. In this article we describe two simple techniquesof using FASTA type sequences and genomic data for the research of genetic markers. In order to apply the functions described below it is necessary to have installed the R language, the *seqRFLP&Maftools*packages, and optionally - the Integrated Development Environment Rstudio.

Body Section:-

SeqRFLPand markers on genetic maps

Simulation and visualization of Restriction Fragment Length Polymorphism (RFLP) patterns resulted from various DNA sequences can be performed using *seqRFLP* R package [1,2].

It includes functions for handling DNA sequences, especially for simulating RFLP patterns based on selected restriction enzymes and creation of so-called *in silico* RFLP genetic maps.The input data consist of FASTA format files and the visualization of the virtual map of simulated DNA digestion can be done with the help of the function *plotenz()*.

```
> library(seqRFLP)
> data(enzdata)
> selected.enzymes=c("BdaI","BstNSI","BtrI","CciI","FatI","PagI")
> plotenz(sequences = dna.seq, enznames = selected.enzymes, enzdata = enzdata, side = FALSE, type =
"RFLP")
```

The *enzdata* allows access to 777 restriction enzymes from which we can select those of interest (*selected.enzymes*) for the research of molecular markers (especially SNP markers) associated with studied sequences (*dna.seq*).

In Fig.1 are presented the RFLP map and the working process for studying polymorphic markers offour virtually fragmented DNA sequences with six restriction enzymes indicated in the *selected.enzymes* vector.

Corresponding Author:- Victor Sitnic

Fig.1:- Studying SNP markers by simulation of restriction enzyme fragmentation [1, 2,4].

SNP molecular markers have a variety of applications in biology and medicine. They are used from species identification to cancer research.

Maftools

Maftools[3] is an R package which in particular facilitates the analysis of oncogenomic data and which incorporates very useful functions for the research of molecular markers. The requested input data are MAF files or even simple TXT files containing the following 9 columns: *Hugo_Symbol, Chromosome, Start_Position, End_Position, Reference_Allele, Tumor_Seq_Allele2, Variant_Classification, Variant_Type, Tumor_Sample_Barcode*.Also for the detection of molecular markers associated with patient survival, are required clinical data which must contain at least 3 columns:*Tumor_Sample_Barcode, Overall_Survival_Status*and*Days_to_last_followup*.

The Tumor_Sample_Barcodecolumn values of the clinical data must correspond to the values of the same column in the genomic data. The reading of data is done by using the function*read.maf()*.

```
> library(maftools)
> input.data = read.maf(maf = genomic.data, clinicalData = clinical.data)
```

The *genomic.data* and *clinical.data* are R dataframe objects and must contain the columns mentioned above. Subsequently, the package allows the use of the *survGroup()* function for detection of potential sets of mutated genes associated with decreased patient survival.

```
>predicted.genes = survGroup(maf = input.data, top = 10, geneSetSize = 2, time = "days_to_last_followup", Status = "Overall_Survival_Status")
```

This function presents an increased interest for the detection of prognostic and risk stratification biomarkers in various pathologies and conditions.

```
> input.data = read.maf(maf = genomic.data, clinicalData = clinical.data)
-Validating
-Silent variants: 13968
-Summarizing
--Possible FLAGS among top ten genes:
  TTN
  MUC16
  SYNE1
-Processing clinical data
-Finished in 11.9s elapsed (11.3s cpu)
> predicted.genes = survGroup(maf = input.data, top = 10, geneSetSize = 2,
 time = "days_to_last_followup", Status = "Overall_Survival_Status")
------------------
genes: 10
geneset size: 2
45 combinations
Looking for clinical data in annoatation slot of MAF..
Removed 120 samples with NA's
Geneset:  TP53,TTN [N= 102 ]
Geneset:  TP53,CSMD3 [N= 52 ]
Geneset:  TP53,FAT1 [N= 51 ]
Geneset:  TP53,CDKN2A [N= 61 ]
```

Fig.2:- Applying the *survGroup()* function to predict the setsof genes associated with decreased patient survival [1,3,4].

The arguments of the function are as follows [3]: *maf* - an MAF object generated by the function *read.maf()*;*top* – top mutated genes; *geneSetSize* – choose desired geneset size; *time* – time column name in the clinical.data object; *Status* - column name containing status of patients in the clinical.data object. Must be logical or numeric (1-deceased, 0 - living).By using *mafSurvGroup()* function it is also possible to plot the survival curves of desired setof genes and wild-type group.

In addition to the above, the R language contains a variety of functions useful for research and identification of genetic markers in various fields of biology and medicine. Its advantage compared to other software platforms is the simplicity of the process of preparation, analysis and visualization of data, sometimes thatbeing performed with only one line of code.

Acknowledgement:-

The articlehas been written within the national project coded 20.80009. 8007.0.

References:-

1. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
2. Qiong Ding and Jinlong Zhang (2012). seqRFLP: Simulation and visualization of restriction enzyme cutting pattern from DNA sequences. R package version 1.0.1. https://CRAN.R-project.org/package=seqRFLP.
3. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. 2018. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Research. http://dx.doi.org/10.1101/gr.239244.118.
4. RStudio Team (2021). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.