SNPannotator package

This package can be used to investigate the functional characteristics of selected SNPs and their vicinity genomic region (up to 500 kb at either sides). Linked SNPs in moderate to high linkage disequilibrium (e.g. r2>0.50) with the corresponding index SNPs will be selected for further analysis. GRCh37 and 38 panels are available. Please visit our website at http://GWASinspector.com for more information and additional datasets.

Data source:

- Variant information is directly fetched from Ensembl database using the REST API server, including:
    - Allele frequency, …
    - Variants in high LD (up to 500 kb at either sides)
    - CADD scores for each allele
    - Phenotype associations
- Gene information is added using an external database. If the file is not provided this information would also be fetched from ENSEMBL website which slightly increases the run time of the job.
    - Gene name and ID are reported for variants that are on a gene
    - Nearest **protein-coding gene** and distances are returned for intergenic variants
- Regulatory features of the region are added using an embedded database.
    - Promoters (regions at the 5' end of genes where transcription factors and RNA polymerase bind to initiate transcription)
    - Promoter flanking regions (transcription factor binding regions that flank the above)
    - Enhancers (regions that bind transcription factors and interact with promoters to stimulate transcription of distant genes)
    - CTCF binding sites (regions that bind CTCF, the insulator protein that demarcates open and closed chromatin)
    - Transcription factor binding sites (sites which bind transcription factors, for which no other role can be determined as yet)
    - Open chromatin regions (regions of spaced out histones, making them accessible to protein interactions)

Supplementary material:

Additional datasets are available from the project website. You can download them to your computer and provide their paths to the main function.

1- Gene name and positions file for the ‘**`geneNames.file`**’ parameter (GRCh37,38)
2- Regulatory features file for the **`‘regulatoryType.file’`** parameter ((GRCh37,38)

Installation & usage

This package depends on the following packages to be installed in R.

- `data.table`
- `httr`
- `jsonlite`
- `xml2`
- `openxlsx`
- `progress`
- `doParallel`
- `foreach`

1- Install the dependency list:

```
install.packages(c("httr","jsonlite","xml2","openxlsx","progress","doParallel))
```

2- Install the package from source file:

```
# Change working directory
setwd('D:/SNPannotator')

# install the package
install.packages("SNPannotator_0.2.0.0.tar.gz", repos = NULL, type='source')
```

3- Run the pipeline on your data:

```
# load the library
library(SNPannotator)

# select server for GRCh38 or GRCh37
# server <- "https://rest.ensembl.org" ### GRCh38
server <- "https://grch37.rest.ensembl.org" ### GRCh37

# select database population for LD calculation
# db <- "1000GENOMES:phase_3:ALL" ### all samples in 1000G study
db <- "1000GENOMES:phase_3:EUR" ### European super population in 1000G study

# create a vector from variant rs numberss
rslist=c('rs16198','rs7469')


# run the pipeline
# the result will be returned as a data frame and also saved as an excel file
# r2: Measure of LD.
# window_size: Window size in kb. The maximum allowed value for the window size is
500 kb. LD is computed for the given variant and all variants that are located within
the specified window.
# LDlist: If set to TRUE, variants in high LD will be found and added to the output.
If FALSE, only the information for variants in rslist will be returned.
# cadd: If set to TRUE the CADD scores will be added to variant information.
# geneNames.file: path the gene information file (*.rds). Default value is NULL and
ENSEMBL website will be checked if no file is provided.
# regulatoryType.file: path the variants regulatory type information file (*.rds).
Default value is NULL and this step will be skipped if no file is provided.

# 1) find variants in high LD with the rslist and fetch information for all
# add cadd score and regulatory type
      output <- annotate(rslist,server,db, 'sampleOutput.xlsx',
             window_size = 500, r2 = .5, LDlist = TRUE, cadd = TRUE,
             geneNames.file='Gene_Names_Ensembl_104_GRCh37.rds',
             regulatoryType.file='homo_sapiens.GRCh37.Regulatory_Build.rds')



# 2) fetch information for the rslist without adding variants in high LD
# add cadd score
       output <- annotate(rslist,server,db, 'sampleOutput.xlsx',
                         LDlist = FALSE, cadd = TRUE)
```

Additional functions:

```
# find variants in high LD with a target variant
       output <- LDlist(rslist,server,db,
                window_size = 500,
                r2 = .5)
```