geneticsCRE

Carl Tony Fakhry and Kourosh Zarringhalam

2017-09-28

geneticsCRE is an R package that performs pathway-based genome-wide association study (PGWAS) to identify statistically significant associations between variants on a gene regulatory pathways and a given phenotype. Unlike Genome-wide association study (GWAS), that seeks to assign statistical significance to associations of variations in single genes to a phenotype, PGWAS accumulates statistical power by examining rare variant along gene-gene interaction pathways. PGWAS uses prior causal information a gene regulatory interactions to infer statistically significant associations between causal pathways and a the phenotype. Given phenotype data with case/control information, **geneticsCRE** computes PGWAS for all valid pathways as identified by the Homo Sapien STRINGdb causal network.

Usage

Processing PGWAS over STRINGdb

geneticsCRE provides simplified functionality for computing PGWAS over STRINGdb. For example, PGWAS can be computed using the following:

```r
library(geneticsCRE)
```

```
## Warning: package 'plyr' was built under R version 3.3.2
```

```r
# Get file of random phenotype data
data_path <- system.file("extdata", "random.phenotype.data", package = "geneticsCRE")

# Compute PGWAS
CRE_Results <- PGWAS(dataset = data_path, nCases = 100, nControls = 100,
                     Signed.PGWAS = FALSE, Decorated.Pvalues = TRUE, threshold = 0.05,
                     K = 10, pathLength = 3, n_permutations = 100, strataF = NA)
```

```
## [1] "Precomputing Scoring Table..."
## [1] "Processing Phenotype dataset..."
## [1] "Processing Network..."
## [1] "Computing PGWAS..."
## [1] "Computing Decorated Pvalues..."
## [1] "Done."
```

`PGWAS` returns a list containing two data frames. The first data frame is `PGWAS.Results` which contains the top `K` paths for each length sorted in increasing order of the p-values. The results are stored in a data frame with the following columns: `SignedPaths` is the column of the top `K` signed paths for each length, `Paths` is the column of the top `K` paths for each length (not including the signs), `Lengths`, `Scores` and `Pvalues` are the length, score and p-value respectively of each path, `Cases` and `Controls` are the number of cases and controls respectively of each path. Even though the signs in `SignedPaths` are reported, since we set `Signed.PGWAS = FALSE` then PGWAS does not take the signs of the path into account when computing the scores of the paths. Since our data is random, we see that none of the p-values are significant if we are to consider a 0.05 significance level.

```r
head(CRE_Results$PGWAS.Results[,c("SignedPaths", "Paths", "Pvalues")])
```

```
##                 SignedPaths           Paths Pvalues
## 11   DOK1 (+) -> DUSP1 (+)   DOK1 -> DUSP1    0.65
## 12     IL4 (+) -> EPOR (+)     IL4 -> EPOR    0.65
## 13 DUSP1 (+) -> PIK3C3 (-) DUSP1 -> PIK3C3    0.65
## 14   DUSP1 (+) -> DOK1 (+)   DUSP1 -> DOK1    0.65
## 15     EPOR (+) -> IL4 (+)     EPOR -> IL4    0.65
## 16 PIK3C3 (+) -> DUSP1 (-) PIK3C3 -> DUSP1    0.65
```

If `Decorated.Pvalues = TRUE`, then the decorated p-values will be computed and the results are stored in a the second data frame `Decorated.Pvalues.Results`. The columns `SignedPaths`, `Paths`, `Lengths`, `Scores`, `Pvalues`, `Cases` and `Controls` in `Decorated.Pvalues.Results` have the same interpretation as in the `PGWAS.Results` data frame. The decorated p-values test whether adding a node to the path is statistically significant. This is done in both directions, going forward from the beginning to the end of the path, and going backwards from the end to the beginning of the path.

```
head(CRE_Results$Decorated.Pvalues.Results[which(CRE_Results$Decorated.Pvalues$Lengths==3)
                                   ,c("Paths", "Subpaths1", "Subpaths2",
                                   "DecoratedPvalues", "Direction")])
```

```
##                     Paths       Subpaths1 Subpaths2 DecoratedPvalues
## 31    TBX21 -> IL4 -> EPOR          TBX21       IL4             0.50
## 32    TBX21 -> IL4 -> EPOR   TBX21 -> IL4      EPOR             0.09
## 33    TBX21 -> IL4 -> EPOR           EPOR       IL4             0.41
## 34    TBX21 -> IL4 -> EPOR    EPOR -> IL4     TBX21             0.50
## 35 TRIB2 -> MAPK14 -> EPOR          TRIB2    MAPK14             1.00
## 36 TRIB2 -> MAPK14 -> EPOR TRIB2 -> MAPK14      EPOR             0.13
##    Direction
## 31   Forward
## 32   Forward
## 33  Backward
## 34  Backward
## 35   Forward
## 36   Forward
```

Processing Signed-PGWAS over STRINGdb

Signed-PGWAS is modified version of PGWAS as it takes the signs of the direction of perturbation into account. It can be called by setting `Signed.PGWAS = TRUE`.

```
# Compute Signed-PGWAS
CRE_Results <- PGWAS(dataset = data_path, nCases = 100, nControls = 100,
                    Signed.PGWAS = TRUE, Decorated.Pvalues = TRUE, threshold = 0.05,
                    K = 10, pathLength = 3, n_permutations = 100, strataF = NA,
                    nthreads = 4)
```

```
## [1] "Precomputing Scoring Table..."
## [1] "Processing Phenotype dataset..."
## [1] "Processing Network..."
## [1] "Computing Signed-PGWAS..."
## [1] "Computing Decorated Pvalues..."
## [1] "Done."
```

Moreover, the decorated p-values can be obtained in a similar way as the unsigned case before:

```
head(CRE_Results$Decorated.Pvalues.Results[which(CRE_Results$Decorated.Pvalues$Lengths==2)
                                   ,c("SignedPaths", "Subpaths1", "Subpaths2",
                                   "DecoratedPvalues")])
```

```
##                 SignedPaths Subpaths1 Subpaths2 DecoratedPvalues
## 11    IHH (+) -> WNT3A (+)       IHH     WNT3A             0.30
## 12    IHH (+) -> WNT3A (+)     WNT3A       IHH             0.26
## 13 MAPKAP1 (+) -> TSC2 (+)   MAPKAP1      TSC2             0.53
## 14 MAPKAP1 (+) -> TSC2 (+)      TSC2   MAPKAP1             0.12
## 15    WNT3A (+) -> IHH (+)     WNT3A       IHH             0.24
## 16    WNT3A (+) -> IHH (+)       IHH     WNT3A             0.29
```

References

[1] Franceschini, A (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. In:'Nucleic Acids Res. 2013 Jan;41(Database issue):D808-15. doi: 10.1093/nar/gks1094. Epub 2012 Nov 29'.