

geneticsCRE

Carl Tony Fakhry and Kouros Zarringhalam

2018-07-26

geneticsCRE is an R package that performs genome-wide association study pathway analysis (GWASPA) to identify statistically significant associations between variants on 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, GWASPA accumulates statistical power by examining rare variant along gene-gene interaction pathways. GWASPA 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 GWASPA for all valid pathways as identified by the Homo Sapien STRINGdb [1] causal network.

Usage

Processing GWASPA over STRINGdb

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

```
library(geneticsCRE)

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

# Compute GWASPA
CRE_Results <- GWASPA(dataset = data_path, nCases = 100, nControls = 100,
                      Signed.GWASPA = FALSE, 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 GWASPA..."
## [1] "Computing Decorated Pvalues..."
## [1] "Done."
```

GWASPA returns a list containing two data frames. The first data frame is **GWASPA.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.GWASPA = FALSE** then GWASPA does not take the signs of the path into account when computing the scores of the paths. Since our data is random, we expect that none of the paths are significant if we are to consider a 0.05 significance level.

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

##               SignedPaths           Paths Pvalues
## 1 MAPKAP1 (+) -> TSC2 (+) MAPKAP1 -> TSC2    0.73
```

```
## 2 TSC2 (+) -> MAPKAP1 (+) TSC2 -> MAPKAP1 0.73
## 3 DUSP1 (+) -> DUT (-) DUSP1 -> DUT 0.73
## 4 IL4 (+) -> EPOR (+) IL4 -> EPOR 0.73
## 5 EPOR (+) -> IL4 (+) EPOR -> IL4 0.73
## 6 IHH (+) -> WNT3A (-) IHH -> WNT3A 0.73
```

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 `GWASPA.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 Direction
## 31 IHH -> WNT3A -> GBX2      IHH      WNT3A           0.24 Forward
## 32 IHH -> WNT3A -> GBX2 IHH -> WNT3A      GBX2           0.43 Forward
## 33 IHH -> WNT3A -> GBX2      GBX2      WNT3A           0.22 Backward
## 34 IHH -> WNT3A -> GBX2 GBX2 -> WNT3A      IHH           0.27 Backward
## 35 IHH -> WNT3A -> WIF1      IHH      WNT3A           0.25 Forward
## 36 IHH -> WNT3A -> WIF1 IHH -> WNT3A      WIF1           0.57 Forward
```

Processing Signed-GWASPA over STRINGdb

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

```
# Compute Signed-GWASPA
CRE_Results <- GWASPA(dataset = data_path, nCases = 100, nControls = 100,
Signed.GWASPA = 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-GWASPA..."
## [1] "Computing Decorated Pvalues..."
## [1] "Done."
```

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

```
##           SignedPaths      Paths Pvalues
## 1 MAPKAP1 (+) -> TSC2 (+) MAPKAP1 -> TSC2 0.71
## 2 EPOR (+) -> IL4 (+) EPOR -> IL4 0.71
## 3 ATM (+) -> DCLRE1C (+) ATM -> DCLRE1C 0.71
## 4 DCLRE1C (+) -> ATM (+) DCLRE1C -> ATM 0.71
## 5 IHH (+) -> WNT3A (+) IHH -> WNT3A 0.71
## 6 IL4 (+) -> EPOR (+) IL4 -> EPOR 0.71
```

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 MAPKAP1 (+) -> TSC2 (+)   MAPKAP1      TSC2              0.56
## 12 MAPKAP1 (+) -> TSC2 (+)      TSC2      MAPKAP1          0.11
## 13      EPOR (+) -> IL4 (+)     EPOR        IL4              0.49
## 14      EPOR (+) -> IL4 (+)     IL4        EPOR              0.19
## 15      ATM (+) -> DCLRE1C (+)   ATM        DCLRE1C          0.24
## 16      ATM (+) -> DCLRE1C (+) DCLRE1C      ATM          0.30
```

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'.