Package ‘PCAmatchR’

April 24, 2020

Title   Match Cases to Controls Based on Genotype Principal Components
Version 0.1.6
Maintainer Derek W. Brown <derek.brown@nih.gov>
Description Matches cases to controls based on genotype principle components (PC).
In order to produce better results, matches are based on the weighted
distance of PCs where the weights are equal to the % variance explained
by that PC. A weighted Mahalanobis distance metric (Kidd et al. (1987)
<DOI:10.1016/0031-3203(87)90066-5>) is used to determine matches.

License  MIT + file LICENSE
URL    https://github.com/machiela-lab/PCAmatchR
BugReports https://github.com/machiela-lab/PCAmatchR/issues
Encoding UTF-8
LazyData true
Depends R (>= 2.10)
RoxygenNote 7.0.2
Suggests optmatch, testthat
NeedsCompilation no
Author Derek W. Brown [aut, cre],
   Mitchel J. Machiela [aut],
   Timothy A. Myers [ctb],
   NCI [cph, fnd]
Repository CRAN
Date/Publication 2020-04-24 11:20:03 UTC

R topics documented:
eigenvalues_1000G ................................. 2
match_maker ........................................ 2
PCs_1000G ......................................... 4
plot_maker ......................................... 5

Index 6
**eigenvalues_1000G**  
*Eigenvalues of 2504 individuals from the 1000 Genome Project*

**Description**
A sample dataset containing the first 20 eigenvalues calculated from 2504 individuals in the Phase 3 data release of the 1000 Genomes Project. The principal component analysis was conducted using PLINK.

**Usage**
eigenvalues_1000G

**Format**
A data frame with 20 rows and 1 variable:
- **eigen_values** calculated eigenvalues

**Source**
Machiela Lab

**Examples**
```r
eigenvalues_1000G
genome_values <- eigenvalues
data <- c(genome_values)$eigen_values
```

**match_maker**  
*Weighted matching of controls to cases using PCA results.*

**Description**
Weighted matching of controls to cases using PCA results.

**Usage**
```r
match_maker(
  PC = NULL,
  eigen_value = NULL,
  data = NULL,
  ids = NULL,
  case_control = NULL,
  num_controls = 1,
```
match_maker

num_PCs = 1000,
exact_match = NULL,
weight_dist = TRUE,
weights = NULL
)

Arguments

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>Individual level principle components.</td>
</tr>
<tr>
<td>eigen_value</td>
<td>Computed eigen value for each PC. Used to calculate the percent variance explained by each PC.</td>
</tr>
<tr>
<td>data</td>
<td>Dataframe containing id and case/control status. Optionally includes covariate data for exact matching.</td>
</tr>
<tr>
<td>ids</td>
<td>The unique id variable contained in both &quot;PC&quot; and &quot;data.&quot;</td>
</tr>
<tr>
<td>case_control</td>
<td>The case control status variable.</td>
</tr>
<tr>
<td>num_controls</td>
<td>The number of controls to match to each case. Default is 1:1 matching.</td>
</tr>
<tr>
<td>num_PCs</td>
<td>The total number of PCs calculated within the PCA. Default is 1000.</td>
</tr>
<tr>
<td>exact_match</td>
<td>Optional variables contained in the dataframe on which to perform exact matching (i.e. sex, race, etc.).</td>
</tr>
<tr>
<td>weight_dist</td>
<td>When set to true, matches are produced based on PC weighted Mahalanobis distance. Default is TRUE.</td>
</tr>
<tr>
<td>weights</td>
<td>Optional user defined weights used to compute the weighted Mahalanobis distance metric.</td>
</tr>
</tbody>
</table>

Value

A list of matches and weights.

Examples

```r
# Create PC data frame by subsetting provided example dataset
cov_data <- as.data.frame(PCs_1000G[,c(15:24)])
# Create eigen values vector using example dataset
eigen_vals <- c(eigenvalues_1000G)eigen_values
# Create Covarite data frame
cov_data <- PCs_1000G[,c(1:4)]
# Generate a case status variable using ESN 1000 Genome population
cov_data$case <- ifelse(cov_data$pop=="ESN", c(1), c(0))
# With 1 to 1 matching
if(!requireNamespace("optmatch", quietly = TRUE)){
mached_makcer(PC = pcs,
eigen_value = eigen_vals,
data = cov_data,
ids = c("sample"),
case_control = c("case"),
num_controles = 1,
num_PCs = dim(cov_data)[1]
)
Principal components of 2504 individuals from the 1000 Genome Project

Description

A sample dataset containing information about population, gender, and the first 20 principal components calculated from 2504 individuals in the Phase 3 data release of the 1000 Genomes Project. The principal component analysis was conducted using PLINK.

Usage

PCs_1000G

Format

A data frame with 2504 rows and 24 variables:

- **sample** sample ID number
- **pop** three letter designation of 1000 Genomes reference population
- **super_pop** three letter designation of 1000 Genomes reference super population
- **gender** gender of individual
- **PC1** principal component 1
- **PC2** principal component 2
- **PC3** principal component 3
- **PC4** principal component 4
- **PC5** principal component 5
- **PC6** principal component 6
- **PC7** principal component 7
- **PC8** principal component 8
- **PC9** principal component 9
- **PC10** principal component 10
- **PC11** principal component 11
- **PC12** principal component 12
- **PC13** principal component 13
- **PC14** principal component 14
- **PC15** principal component 15
- **PC16** principal component 16
- **PC17** principal component 17
- **PC18** principal component 18
- **PC19** principal component 19
- **PC20** principal component 20
**plot_maker**

*Function to plot matches from match_maker output*

---

**Description**

Function to plot matches from match_maker output

**Usage**

```r
plot_maker(
  data = NULL,
  x_var = NULL,
  y_var = NULL,
  case_control = NULL,
  line = T
)
```

**Arguments**

- `data`: match_maker output
- `x_var`: Principal component 1
- `y_var`: Principal component 2
- `case_control`: Case or control status
- `line`: draw line

**Value**

None

**Examples**

```r
plot_maker(data=match_maker_output,
  x_var="PC1",
  y_var="PC2",
  case_control="case",
  line=T)
```
Index

*Topic datasets
  eigenvalues_1000G, 2
  PCs_1000G, 4

eigenvalues_1000G, 2
match_maker, 2
PCs_1000G, 4
plot_maker, 5