

# Package ‘EBPRS’

October 12, 2022

**Type** Package

**Title** Derive Polygenic Risk Score Based on Empirical Bayes Theory

**Version** 2.1.0

**Author** Shuang Song [aut, cre], Wei Jiang [aut], Lin Hou [aut] and Hongyu Zhao [aut]

**Maintainer** Shuang Song <song-s19@mails.tsinghua.edu.cn>

**Description** EB-PRS is a novel method that leverages information for effect sizes across all the markers to improve the prediction accuracy. No parameter tuning is needed in the method, and no external information is needed. This R-package provides the calculation of polygenic risk scores from the given training summary statistics and testing data. We can use EB-PRS to extract main information, estimate Empirical Bayes parameters, derive polygenic risk scores for each individual in testing data, and evaluate the PRS according to AUC and predictive  $r^2$ . See Song et al. (2020) <[doi:10.1371/journal.pcbi.1007565](https://doi.org/10.1371/journal.pcbi.1007565)> for a detailed presentation of the method.

**License** GPL-3

**Depends** R (>= 3.5.0), ROCR, methods, BEDMatrix, data.table

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 6.1.1

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2020-08-26 05:40:08 UTC

## R topics documented:

EBPRS	2
EBPRSpackage	3
read_plink	4
traindat	5
validate	6
<b>Index</b>	<b>7</b>

EBPRS

*Main function***Description**

Clean the dataset, extract information from raw data and calculate effect sizes. (Please notice that there are some requirements for the training and testing datasets.)

**Usage**

```
EBPRS(train, test, N1, N0, robust = T)
```

**Arguments**

<code>train</code>	training dataset
<code>test</code>	testing dataset (list) including fam, bed, bim, which can be generated from function <code>read_plink</code> in our package. If <code>missing(test)=T</code> , the function will use all SNPs in training dataset by default.
<code>N1</code>	case number
<code>N0</code>	control number
<code>robust</code>	T/F, indicator that whether robust estimation is needed. The function will run faster when robust is set to F. The default is T.

**Details**

The raw training data should be a data.frame including A1, A2, OR, P, SNP (NOTE that the colnames should be exactly consistent with the above).

The SNP column (rsid) is used for indexing.

An example training dataset can be acquired using `data("traindat")`

"test" file can be generated from `read_plink("path_to_test_plink_bfile")`

`test` is a list, which has `test$fam` (6 columns with information on samples), `test$bim` (6 columns with information on SNPs), `test$bed` (genotypes matrix 0, 1, 2)

Note that in real data, we usually use  $\beta_0 = m/20$  as the default setting for the EM algorithm, which is accurate enough in most cases and will have little influence on the prediction performance. If more accurate parameter estimation is required, we provide a robust estimation (by setting `robust=T`), integrating our data-driven bootstrap-based parameter tuning method. This can derive the best parameter for robust estimation, while more time is needed.

**Value**

A list containing data.frame (result): combining the summary statistics and estimated effect sizes (eff)

estimated effect sizes (eff)

estimated mu (muHat)

estimated sigma2 (sigmaHat2)  
estimated proportion of non-associated SNPs (pi0)  
estimated variance of effect sizes of associated SNPs (sigma02)  
If the test file is provided the results also include:  
predictive r2 (r2)  
AUC (AUC)  
estimated polygenic risk score (S)

**Author(s)**

Shuang Song, Wei Jiang, Lin Hou and Hongyu Zhao

**References**

Song S, Jiang W, Hou L, Zhao H (2020) Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. *PLoS Comput Biol* 16(2): e1007565. <https://doi.org/10.1371/journal.pcbi.1007565>

---

EBPRSpackage

*Description of the package*

---

**Description**

Description of the package. This is the 2.0.3 version.

**Usage**

EBPRSpackage()

**Details**

EB-PRS is a novel method that leverages information for effect sizes across all the markers to improve the prediction accuracy. No parameter tuning is needed in the method, and no external information is needed. This R-package provides the calculation of polygenic risk scores from the given training summary statistics and test data. We can use EB-PRS to extract main information, estimate Empirical Bayes parameters, derive polygenic risk scores for each individual in test data, and evaluate the PRS according to AUC and predictive r2.

Package: EBPRS  
Type: Package  
Date: 2019-12  
Version: 2.1.0

The package contains three main functions for users, `read_plink`, `EBPRS`, and `validate`.

1. `read_plink`. This function is used to read plink bfiles into R and reformat to suit the input of function `EBPRS()`.

2. `EBPRS`. This function integrate three parts: (1) merge the train and test (if have) data, (2) estimate effectsize (3) generate polygenic risk scores (if test data provided.)

There is a strict requirement for the format of input, which is detailedly illustrated in details in function `EBPRS()`. The training summary statistics are necessary. The test data can either be included in the input or not. If test data are provided. The function will first merge the data, as well as generate scores for each person in the result. Users could first use the function `read_plink()` implemented in our package to read plink files into R.

3. `validate`. We use this to validate the performance of the PRS.

4. `data("traindat")` for the example training dataset.

A complete pipeline can be:

```
train <- fread('trainpath') (pay attention to the format, detailed in EBPRS())
```

```
test <- read_plink('testpath') (path to the plink bfile without extensions)
```

```
result <- EBPRS(train=traindat, test=plinkfile, N1, N0)
```

```
validate(result$S, truey)
```

or

```
train <- fread('trainpath') (pay attention to the format)
```

```
result <- EBPRS(train=traindat, N1, N0) (will only provide estimated effect sizes)
```

### Author(s)

Shuang Song, Wei Jiang, Lin Hou and Hongyu Zhao

### References

Song S, Jiang W, Hou L, Zhao H (2020) Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. *PLoS Comput Biol* 16(2): e1007565. <https://doi.org/10.1371/journal.pcbi.1007565>

### See Also

[EBPRS](#), [validate](#),

---

read\_plink

*Read plink bfiles to R and reformat*

---

### Description

To read plink files into R and transfer the files to the format we use in the `EBPRS()` function.

### Usage

```
read_plink(path)
```

**Arguments**

path                    path to the test files (plink bfiles, without extension)

**Details**

The input should not include the extension. For example, the test files are AA.bed, AA.bim and AA.fam, then the input should be 'AA' instead of 'AA.bed'.

**Author(s)**

Shuang Song, Wei Jiang, Lin Hou and Hongyu Zhao

**References**

Song S, Jiang W, Hou L, Zhao H (2020) Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. PLoS Comput Biol 16(2): e1007565. <https://doi.org/10.1371/journal.pcbi.1007565>

**See Also**

[EBPRS](#)

---

traindat	<i>Example data for training set</i>
----------	--------------------------------------

---

**Description**

Summary statistics simulated in the manuscript Leveraging effect size distributions to improve polygenic risk scores derived from genome-wide association studies. Data from a QTL experiment on gravitropism in

**Usage**

```
data("traindat")
```

**Format**

data.frame

**References**

Song S, Jiang W, Hou L, Zhao H (2020) Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. PLoS Comput Biol 16(2): e1007565. <https://doi.org/10.1371/journal.pcbi.1007565>

**Examples**

```
data("traindat")
## Not run:
result=EBPRS(train=traindat, N1=364, N0=2063)
## End(Not run)
```

---

validate

*Validate the performance of EBPRS*

---

**Description**

Provide the performance evaluated by predictive r2 and AUC.

**Usage**

```
validate(score, truey)
```

**Arguments**

score	polygenic score generated by 'EBPRS'
truey	the true phenotype (the 6th column in the fam file)

**Author(s)**

Shuang Song, Wei Jiang, Lin Hou and Hongyu Zhao

**References**

Song S, Jiang W, Hou L, Zhao H (2020) Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. *PLoS Comput Biol* 16(2): e1007565. <https://doi.org/10.1371/journal.pcbi.1007565>

**See Also**

[EBPRS](#)

**Examples**

```
validate(score=rnorm(20,0,1), truey=sample(0:1,20,replace=TRUE))
```

# Index

## \* datasets

traindat, 5

EBPRS, 2, 4–6

EBPRSPackage, 3

read\_plink, 4

traindat, 5

validate, 4, 6