Tutorial

Szu-Ping,Chen

In this tutorial, we will explore the step-by-step process of using EHPGS package to evaluate hybrid performance. The dataset used for demonstration is the pumpkin dataset published by Wu et al. (2019).

Data requirements

Before evaluation, there are three, please ensure that you have the following data available: 1. training set's genotype: It is a numeric matrix representing the training set, where each row represents an individual and each column represents a single SNP. 2. training set's phenotype: A numeric vector of phenotype of training set. 3. parental lines' genotype: It is a numeric matrix representing the parental lines, where each row represents an individual and each column represents a single SNP.

```
print(train.geno[1:5,1:5])
```

```
##           S1_129359 S1_132420 S1_174510 S1_175873 S1_182575
## P007:P279         1         1         0         0         0
## P008:P252         1         1         1         1         1
## P009:P131         0         0         1         1         1
## P010:P012         1         1         0         0         0
## P010:P261         1         1         0         0         0
```

```
print(train.pheno[1:5,])
```

```
##       P1   P2    Hybrid F1.weight
## 113 P007 P279 P007:P279 2.4638168
## 235 P008 P252 P008:P252 2.7248522
## 328 P009 P131 P009:P131 2.1713029
## 422 P010 P012 P010:P012 0.4952866
## 518 P010 P261 P010:P261 0.9952866
```

```
print(parent.geno[1:5,1:5])
```

```
##      S1_129359 S1_132420 S1_174510 S1_175873 S1_182575
## P007         1         1        -1        -1        -1
## P008         1         1         1         1         1
## P009        -1        -1         1         1         1
## P010         1         1        -1        -1        -1
## P011         1         1        -1        -1        -1
```

Evaluation hybrid performance

Please select the desired prediction method and configure the relevant parameters to proceed. If the training set includes hybrids, kindly provide the names of the hybrids in the following format: two parents separated by a colon (e.g., A:B) and run EHPGS; then we can get the output of best hybrids and parental lines.

```
parent.geno = parent.geno[1:10,]
hybrid.name = combn(row.names(parent.geno), 2, paste, collapse = ":")
```

```
tr.test = intersect(row.names(train.geno),combn(row.names(parent.geno), 2, paste, collapse = ":"))
result = EHPGS(num.F1=15,num.P=5,train.pheno=train.pheno$F1.weight,train.geno = train.geno,parent.geno =
```

```
##
## #chain=1
## iter=.0.1.2.3.4.5.6.7.8.9.10.11.12.13.14.15.16.17.18.19.20.21.22.23.24.25.26.27.28.29.30.31.32.33.34
```

```
print(result)
```

```
## $F1.out
##             GEBV    SCA     MPH    BPH
## P007:P008 2.681   0.256   0.256   0.207
## P008:P027 2.397   0.244   0.244  -0.116
## P007:P009 2.380   0.411   0.411  -0.061
## P008:P009 2.285   0.273   0.273  -0.248
## P007:P027 2.176   0.066   0.066  -0.246
## P009:P027 2.035   0.338   0.338   0.178
## P007:P011 1.971   0.121   0.121  -0.488
## P007:P013 1.895   0.317   0.317  -0.603
## P008:P011 1.881  -0.012  -0.012  -0.670
## P008:P014 1.867   0.192   0.192  -0.715
## P008:P013 1.864   0.244   0.244  -0.725
## P007:P010 1.774   0.147   0.147  -0.717
## P007:P026 1.763   0.238   0.238  -0.743
## P007:P014 1.758   0.125   0.125  -0.732
## P011:P027 1.757   0.179   0.179  -0.118
##
## $P.out
##         GEBV     GCA
## P008 2.570   0.511
## P007 2.473   0.462
## P027 1.849   0.150
## P009 1.528  -0.010
## P011 1.256  -0.146
##
## $V.out
##       mu     vA     vD     vE
## 1 1.548 0.808 0.235 0.315
```

The result shows the prediction value of the top 15 hybrids and 5 parental lines.