

# Estimating Gene-Specific Phenotypes with **gespeR**

Fabian Schmich

March 17, 2017

## Contents

### 1 The **gespeR** Model

This package provides algorithms for deconvoluting off-target confounded phenotypes from RNA interference screens. The package uses (predicted) siRNA-to-gene target relations in a regularised linear regression model, in order to infer individual gene-specific phenotype (GSP) contributions. The observed siRNA-specific phenotypes (SSPs) for reagent  $i = 1, \dots, n$  as the weighted linear sum of GSPs of all targeted genes  $j = 1, \dots, p$

$$Y_i = x_{i1}\beta_1 + \dots + x_{ip}\beta_p + \epsilon_i, \quad (1)$$

where  $x_{ij}$  represents the strength of knockdown of reagent  $i$  on gene  $j$ ,  $\beta_j$  corresponds to the GSP of gene  $j$  and  $\epsilon_i$  is the error term for SSP  $i$ . The linear regression model is fit using elastic net regularization:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \left( y_i - \sum_{j=1}^p x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^p \left( \alpha\beta_j^2 + (1-\alpha)|\beta_j| \right) \right\}. \quad (2)$$

Here  $\lambda$  determines the amount of regularization and  $\alpha$  is the mixing parameter between the ridge and lasso penalty with  $0 \leq \alpha \leq 1$ . The elastic net penalty selects variables like the lasso and shrinks together the coefficients of correlated predictors like ridge. This allows for a sparse solution of nonzero GSPs, while retaining simultaneous selection of genes with similar RNAi reagent binding patterns in their respective 3' UTRs. For more information and for citing the **gespeR** package please refer to:

Schmich F (2015). *gespeR: Gene-Specific Phenotype Estimator*. R package version 1.6.1, <http://www.cbg.ethz.ch/software/gespeR>.

### 2 Working Example

In this example, we first load simulated phenotypic readout and siRNA-to-gene target relations. The toy data consists of four screens (A, B, C, D) of 1,000 siRNAs and a limited gene universe of 1,500 genes. Detailed description of how the data was simulated can be accessed using `?simData`. First, we load the package:

```
library(gespeR)
```

Now the phenotypes and target relations can be initialised using the `Phenotypes` and `TargetRelations` commands. First, we load the four phenotypes vectors:

```

phenos <- lapply(LETTERS[1:4], function(x) {
  sprintf("Phenotypes_screen_%s.txt", x)
})
phenos <- lapply(phenos, function(x) {
  Phenotypes(system.file("extdata", x, package="gesper"),
    type = "SSP",
    col.id = 1,
    col.score = 2)
})
show(phenos[[1]])

## 1000 SSP Phenotypes
##
## # A tibble: 1,000  2
##       ID      Scores
##   <fctr>    <dbl>
## 1 siRNAID_0001 -0.93028719
## 2 siRNAID_0002 -1.12820384
## 3 siRNAID_0003 -1.05265043
## 4 siRNAID_0004  0.80792721
## 5 siRNAID_0005 -1.41533349
## 6 siRNAID_0006  1.64265769
## 7 siRNAID_0007 -0.15733945
## 8 siRNAID_0008  0.74758974
## 9 siRNAID_0009 -0.95904664
## 10 siRNAID_0010 -0.04401824
## # ... with 990 more rows

```

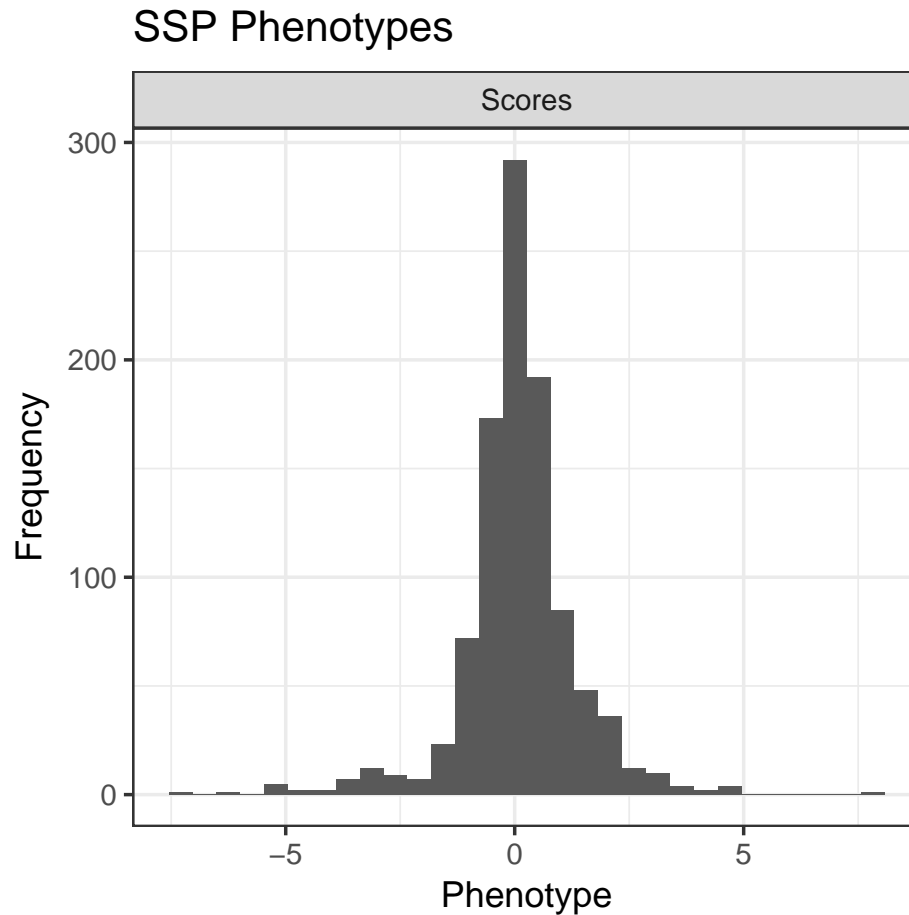
A visual representation of the phenotypes can be obtained with the `plot` method:

```

plot(phenos[[1]])

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

```



Now, we load the target relations for all four screens using the constructor of the `TargetRelations` class:

```
tr <- lapply(LETTERS[1:4], function(x) {
  sprintf("TR_screen_%s.rds", x)
})
tr <- lapply(tr, function(x) {
  TargetRelations(system.file("extdata", x, package="gesperR"))
})
show(tr[[2]])
```

```
## 1000 x 1500 siRNA-to-gene relations.
## 10 x 5 sparse Matrix of class "dgCMatrix"
##           colnames
## rownames  geneID_0001 geneID_0002 geneID_0003 geneID_0004 geneID_0005
## siRNAID_0001      .      .      .      .      .
## siRNAID_0002      .      .      .      .      .
## siRNAID_0003      . 0.4385757      .      .      .
## siRNAID_0004      .      .      .      .      .
## siRNAID_0005      .      .      .      .      .
## siRNAID_0006      .      .      .      .      .
## siRNAID_0007      .      .      .      .      .
## siRNAID_0008      .      .      .      .      .
## siRNAID_0009      .      .      .      .      .
## siRNAID_0010      .      .      .      .      .
## ...
```

For large data sets, e.g. genome-wide screens, target relations objects can become very big and the user may not want to keep all values in the RAM. For this purpose, we can use the `unloadValues` method. In this example, we write the values to a temp-file, i.e. not the original source file, which may be required, when we do not want to overwrite existing data, after, for instance, subsetting the target relations object.

```
# Size of object with loaded values
format(object.size(tr[[1]]), units = "Kb")

## [1] "678.5 Kb"

tempfile <- paste(tempfile(pattern = "file", tmpdir = tempdir()), ".rds", sep="")
tr[[1]] <- unloadValues(tr[[1]], writeValues = TRUE, path = tempfile)

# Size of object after unloading
format(object.size(tr[[1]]), units = "Kb")

## [1] "159.1 Kb"

# Reload values
tr[[1]] <- loadValues(tr[[1]])
```

In order to obtain deconvoluted gene-specific phenotypes (GSPs), we fit four models on the four separate data sets using cross validation by setting `mode = "cv"`. We set the elastic net mixing parameter to 0.5 and use only one core in this example:

```
res.cv <- lapply(1:length(phenos), function(i) {
  gesper(phenotypes = phenos[[i]],
    target.relations = tr[[i]],
    mode = "cv",
    alpha = 0.5,
    ncores = 1)
})
```

The `ssp` and `gsp` methods can be used to obtain SSP and GSP scores from a `gesper` object:

```
ssp(res.cv[[1]])

## 1000 SSP Phenotypes
##
## # A tibble: 1,000  2
##       ID      Scores
##   <fctr>    <dbl>
## 1 siRNAID_0001 -0.93028719
## 2 siRNAID_0002 -1.12820384
## 3 siRNAID_0003 -1.05265043
## 4 siRNAID_0004  0.80792721
## 5 siRNAID_0005 -1.41533349
## 6 siRNAID_0006  1.64265769
## 7 siRNAID_0007 -0.15733945
## 8 siRNAID_0008  0.74758974
## 9 siRNAID_0009 -0.95904664
## 10 siRNAID_0010 -0.04401824
## # ... with 990 more rows

gsp(res.cv[[1]])
```

```
## 1500 GSP Phenotypes
##
## # A tibble: 1,500  2
##       ID Scores
## *    <fctr> <dbl>
## 1 geneID_0001    NA
## 2 geneID_0002    NA
## 3 geneID_0003    NA
## 4 geneID_0004    NA
## 5 geneID_0005    NA
## 6 geneID_0006    NA
## 7 geneID_0007    NA
## 8 geneID_0008    NA
## 9 geneID_0009    NA
## 10 geneID_0010    NA
## # ... with 1,490 more rows

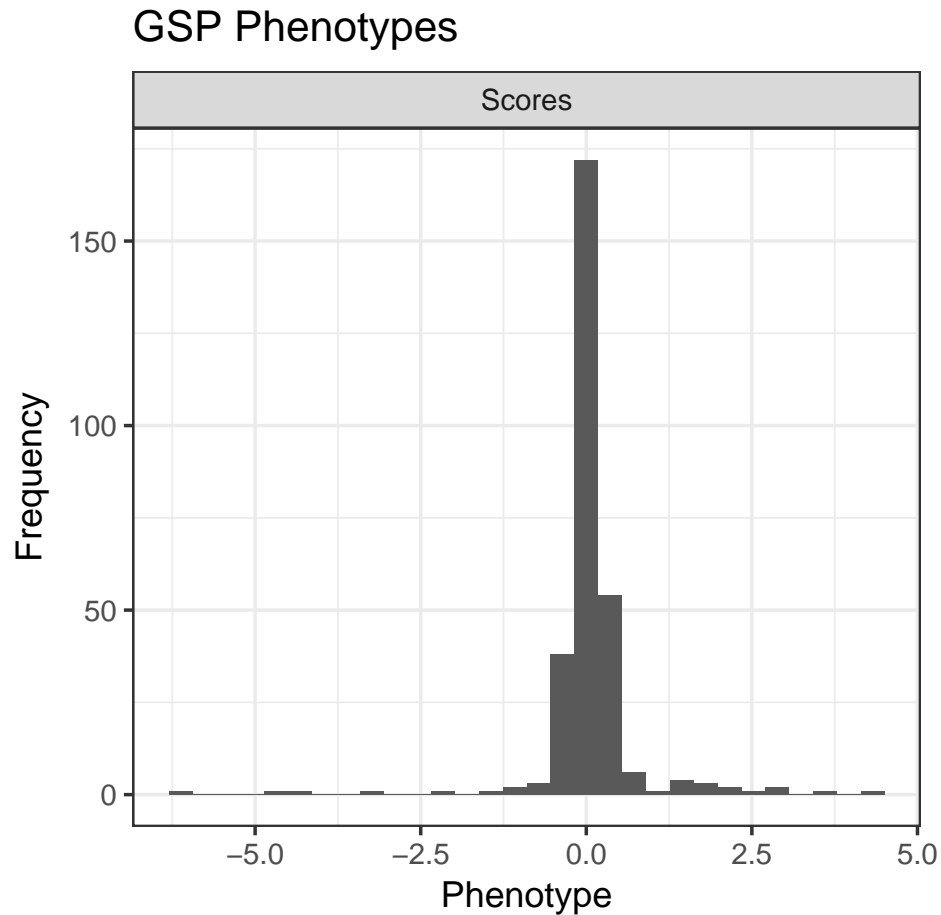
head(scores(res.cv[[1]]))

## # A tibble: 6  2
##       ID Scores
##    <fctr> <dbl>
## 1 geneID_0001    NA
## 2 geneID_0002    NA
## 3 geneID_0003    NA
## 4 geneID_0004    NA
## 5 geneID_0005    NA
## 6 geneID_0006    NA
```

The fitted models can also be visualised using the `plot` method:

```
plot(res.cv[[1]])

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## Warning: Removed 1204 rows containing non-finite values (stat_bin).
```

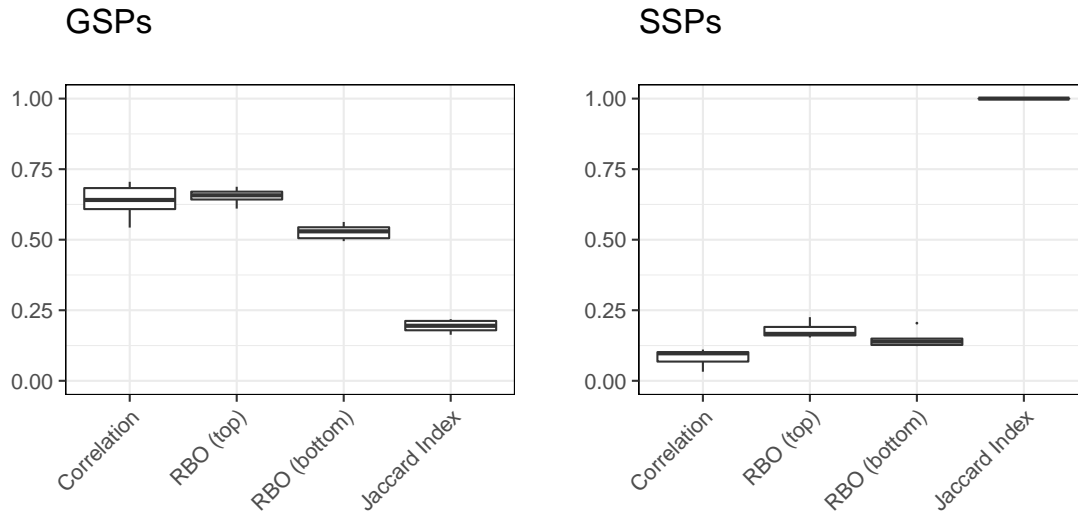


The `concordance` method can be used to compute the concordance between ranked lists of phenotypes. Here we compute concordance between all pairs of GSPs, as well as between all pairs of SSPs, from all four data sets:

```
conc.gsp <- concordance(lapply(res.cv, gsp))  
conc.ssp <- concordance(lapply(res.cv, ssp))
```

We can visualise the `concordance` objects using the `plot` method:

```
plot(conc.gsp) + ggtitle("GSPs\n")  
plot(conc.ssp) + ggtitle("SSPs\n")
```



### 3 sessionInfo()

- R version 3.3.3 (2017-03-06), x86\_64-apple-darwin13.4.0
- Locale: C/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: gesper 1.6.1, ggplot2 2.2.1, knitr 1.15.1
- Loaded via a namespace (and not attached): AnnotationDbi 1.36.2, Biobase 2.34.0, BiocGenerics 0.20.0, BiocInstaller 1.24.0, Category 2.40.0, DBI 0.6, DEoptimR 1.0-8, GSEABase 1.36.0, IRanges 2.8.2, MASS 7.3-45, Matrix 1.2-8, R6 2.2.0, RBGL 1.50.0, RColorBrewer 1.1-2, RCurl 1.95-4.8, RSQLite 1.1-2, Rcpp 0.12.9, S4Vectors 0.12.2, XML 3.98-1.5, affy 1.52.0, affyio 1.44.0, annotate 1.52.1, assertthat 0.1, biomaRt 2.30.0, bitops 1.0-6, cellHTS2 2.38.0, cluster 2.0.6, codetools 0.2-15, colorspace 1.3-2, digest 0.6.12, doParallel 1.0.10, dplyr 0.5.0, evaluate 0.10, foreach 1.4.3, genefilter 1.56.0, glmnet 2.0-5, graph 1.52.0, grid 3.3.3, gtable 0.2.0, highr 0.6, iterators 1.0.8, labeling 0.3, lattice 0.20-34, lazyeval 0.2.0, limma 3.30.13, magrittr 1.5, memoise 1.0.0, munsell 0.4.3, mvtnorm 1.0-6, parallel 3.3.3, pcaPP 1.9-61, plyr 1.8.4, prada 1.50.0, preprocessCore 1.36.0, reshape2 1.4.2, robustbase 0.92-7, rrcov 1.4-3, scales 0.4.1, splines 3.3.3, stats4 3.3.3, stringi 1.1.2, stringr 1.2.0, survival 2.41-2, tibble 1.2, tools 3.3.3, vsn 3.42.3, xtable 1.8-2, zlibbioc 1.20.0