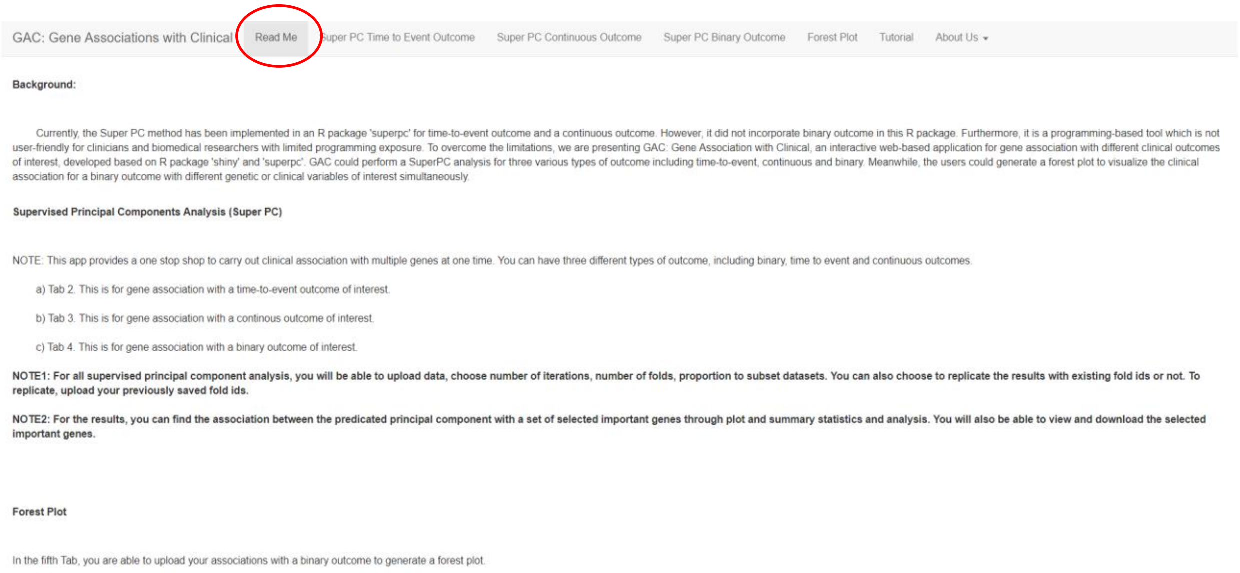
GAC Tutorial

Tab 1: Read Me



GAC: Gene Associations with Clinical | Read Me | Super PC Time to Event Outcome | Super PC Continuous Outcome | Super PC Binary Outcome | Forest Plot | Tutorial | About Us ▾

Background:

Currently, the Super PC method has been implemented in an R package 'superpc' for time-to-event outcome and a continuous outcome. However, it did not incorporate binary outcome in this R package. Furthermore, it is a programming-based tool which is not user-friendly for clinicians and biomedical researchers with limited programming exposure. To overcome the limitations, we are presenting GAC: Gene Association with Clinical, an interactive web-based application for gene association with different clinical outcomes of interest, developed based on R package 'shiny' and 'superpc'. GAC could perform a SuperPC analysis for three various types of outcome including time-to-event, continuous and binary. Meanwhile, the users could generate a forest plot to visualize the clinical association for a binary outcome with different genetic or clinical variables of interest simultaneously.

Supervised Principal Components Analysis (Super PC)

NOTE: This app provides a one stop shop to carry out clinical association with multiple genes at one time. You can have three different types of outcome, including binary, time to event and continuous outcomes.

a) Tab 2. This is for gene association with a time-to-event outcome of interest.

b) Tab 3. This is for gene association with a continous outcome of interest.

c) Tab 4. This is for gene association with a binary outcome of interest.

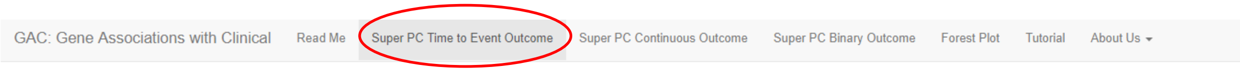
NOTE1: For all supervised principal component analysis, you will be able to upload data, choose number of iterations, number of folds, proportion to subset datasets. You can also choose to replicate the results with existing fold ids or not. To replicate, upload your previously saved fold ids.

NOTE2: For the results, you can find the association between the predicated principal component with a set of selected important genes through plot and summary statistics and analysis. You will also be able to view and download the selected important genes.

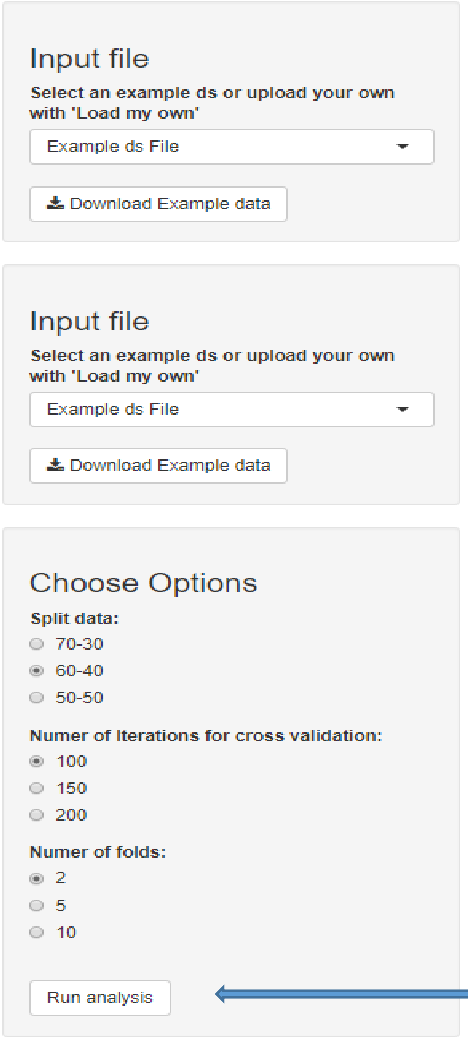
Forest Plot

In the fifth Tab, you are able to upload your associations with a binary outcome to generate a forest plot.

Tab 2: Super PC Time to Event Outcome



GAC: Gene Associations with Clinical | Read Me | Super PC Time to Event Outcome | Super PC Continuous Outcome | Super PC Binary Outcome | Forest Plot | Tutorial | About Us ▾



Input file

Select an example ds or upload your own with 'Load my own'

Example ds File ▾

Download Example data

Step 1:

Select example 'Time to event' clinical dataset or upload your own.

To view example data and format, use download button to view .csv file.

Input file

Select an example ds or upload your own with 'Load my own'

Example ds File ▾

Download Example data

Step 2:

Select example 'expression' dataset or upload your own. The patients should be ordered in same order and should be exact same.

To view example data and format, use download button to view .csv file.

Choose Options

Split data:

○ 70-30

● 60-40

○ 50-50

Numer of Iterations for cross validation:

● 100

○ 150

○ 200

Numer of folds:

● 2

○ 5

○ 10

Run analysis

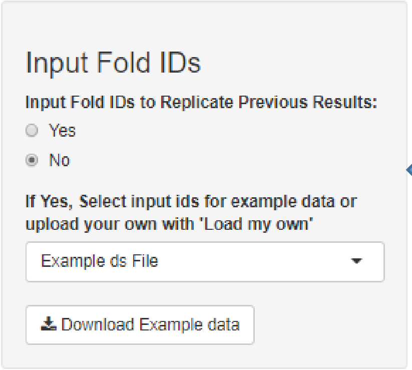
Step 3:

Choose data split into training and validation. Default uses 60-40 i.e. data is split into 40 % for training and remaining 60% for validation. You can also choose the number of interactions. For example purposes a smaller number 100 is used, but greater number of iterations are preferred.

Also, select number of folds for cross validation. Default is 2.

Step 4:

Hit button to run analysis after each change in option.



Input Fold IDs

Input Fold IDs to Replicate Previous Results:

○ Yes

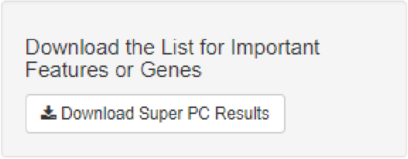
● No

If Yes, Select input ids for example data or upload your own with 'Load my own'

Example ds File

Step 5 (Optional):

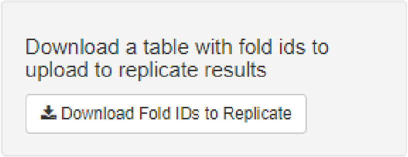
The user can replicate previously generated results by uploading fold-ids (download available below). When running analysis for first time, leave option to 'No'.



Download the List for Important Features or Genes

Step 6 (Optional):

The user can download the tabular results in .csv file.



Download a table with fold ids to upload to replicate results

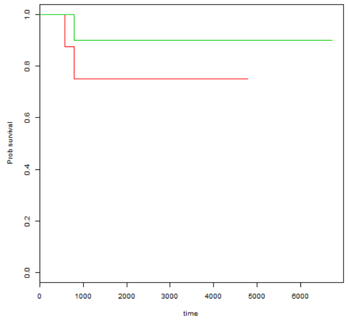
Step 7 (Optional):

The user can download the fold-ids in csv format to replicate results at a later time. These can be uploaded in the input fold id section.

Association between Continuous Predicted Principal Component with Time to Event Outcome

```
The p-value for the first Principal component is = 0.215364005988728.
```

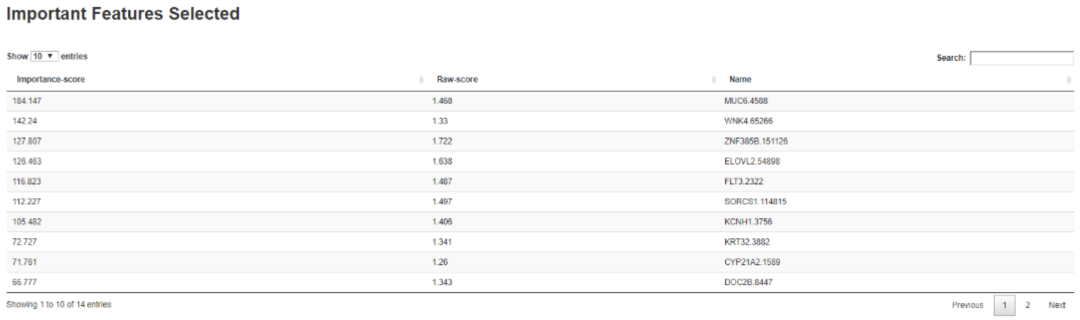
KM Plot for Discrete Predicted Principal Component (Cut by Median)



Main panel shows the results from the super PC analysis.

The p-value from 1st supervised PC is reported.

A discrete (categorical) predictor is created by cutting the predictor at its median to form two groups for the Kaplan-Meier analysis.



Important Features Selected

Show 10 ▾ entries

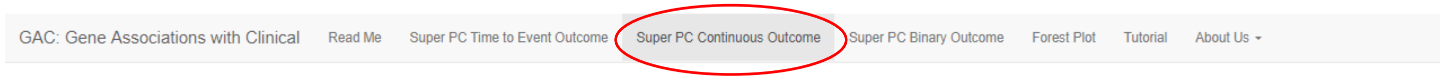
Search:

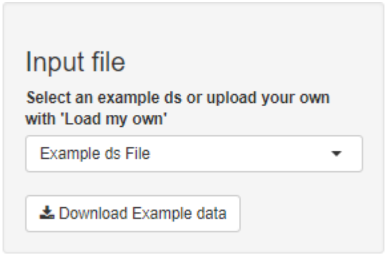
Importance-score	Raw-score	Name
184.147	1.468	MUC6.4588
142.24	1.33	WNK4.65266
127.807	1.722	ZNF385B.151126
126.463	1.638	ELOVL2.54898
116.823	1.487	FLT3.2322
112.227	1.497	SORCS1.114815
105.482	1.406	KCNH1.3756
72.727	1.341	KRT32.3882
71.761	1.26	CYP21A2.1589
66.777	1.343	DOC2B.8447

Showing 1 to 10 of 14 entries

List of all significant genes, in order of decreasing importance score are reported

Tab 3: Super PC Continuous Outcome


GAC: Gene Associations with Clinical
Read Me
Super PC Time to Event Outcome
Super PC Continuous Outcome
Super PC Binary Outcome
Forest Plot
Tutorial
About Us


Input file

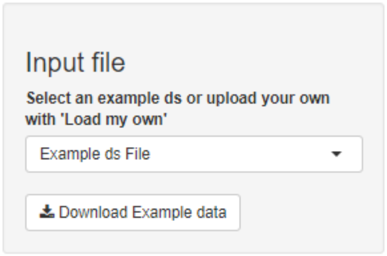
Select an example ds or upload your own with 'Load my own'

Example ds File

Step 1:

Select example ‘continuous’ outcome dataset or upload your own.

To view example data and format, use download button to view .csv file.


Input file

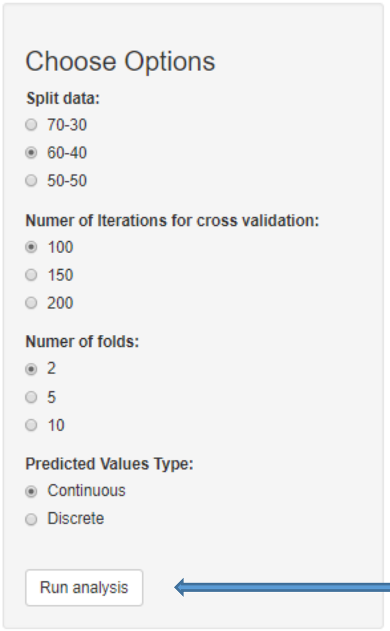
Select an example ds or upload your own with 'Load my own'

Example ds File

Step 2:

Select example ‘expression’ dataset or upload your own. The patients should be ordered in same order and should be exact same.

To view example data and format, use download button to view .csv file.


Choose Options

Split data:

○ 70-30
● 60-40
○ 50-50

Numer of Iterations for cross validation:

● 100
○ 150
○ 200

Numer of folds:

● 2
○ 5
○ 10

Predicted Values Type:

● Continuous
○ Discrete

Step 3:

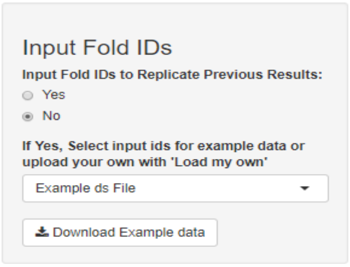
Choose data split into training and validation. Default uses 60-40 i.e. data is split into 40 % for training and remaining 60% for validation. You can also choose the number of interactions. For example purposes a smaller number 100 is used, but greater number of iterations are preferred.

Also, select number of folds for cross validation. Default is 2.

Additionally, choose to display predicted values as scatter plots (for continuous predictors with Pearson correlation’s) or boxplot (for binary groups cut off at median with t-test).

Step 4:

Hit button to run analysis after each change in option.


Input Fold IDs

Input Fold IDs to Replicate Previous Results:

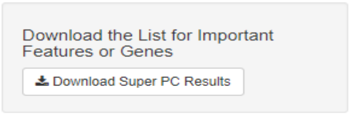
○ Yes
● No

If Yes, Select input ids for example data or upload your own with 'Load my own'

Example ds File

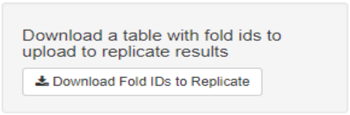
Step 5 (Optional):

The user can replicate previously generated results by uploading fold-ids (download available below). When running analysis for first time, leave option to ‘No’.


Download the List for Important Features or Genes

Step 6 (Optional):

The user can download the tabular results in .csv file.


Download a table with fold ids to upload to replicate results

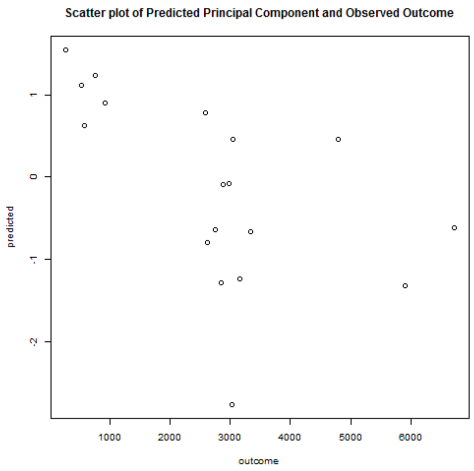
Step 7 (Optional):

The user can download the fold-ids in csv format to replicate results at a later time. These can be uploaded in the input fold id section.

Association between Continuous Predicted Principal Component with Continuous Outcome

```
The p-value for the first Principal component is = 0.0651810699008778.
```

Plot and Tests for Predicted Principal Component



```
        Pearson's product-moment correlation

data:  outcome and predicted
t = -2.7285, df = 16, p-value = 0.01488
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8157644 -0.1311425
sample estimates:
       cor
-0.5635107
```

Main panel shows the results from the super PC analysis.

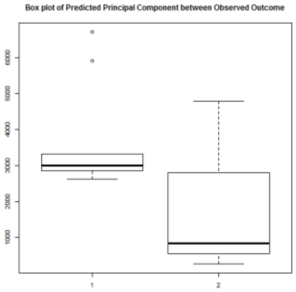
The p-value from 1st supervised PC is reported.

Scatter plot showing correlation of continuous predictor with outcome using Pearson correlation is reported. Discrete predictors are created by cutting the predictor at its median to form two groups and reported as boxplots.

Predicted Values Type:

○ Continuous

● Discrete



```
       Welch Two Sample t-test

data:  dat$outcome by dat$pred.discrete
t = 2.6516, df = 14.269, p-value = 0.01873
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  373.9631 3510.3369
sample estimates:
mean in group 1 mean in group 2
       3620.90         1678.75
```

Important Features Selected

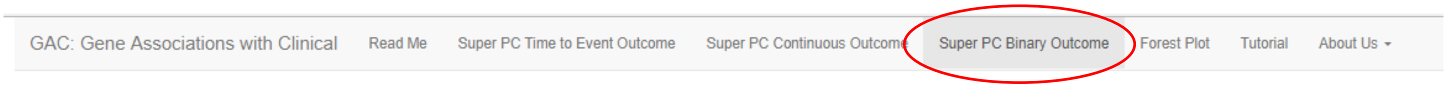
Show 10 ▼ entries   Search:

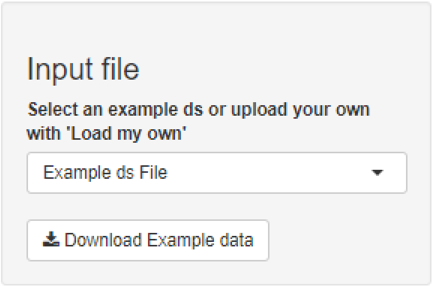
Importance-score	Raw-score	Name
-87.439	-0.707	UGT2B11.10720
-41.829	0.699	PYDC1.260434
-41.136	-0.863	UGT2B28.54490
39.097	1.244	XK.7504
-38.889	-0.757	FCN2.2220
38.537	0.761	TGFA.7039
38.37	0.835	DPYSL5.56896
38.306	1.112	GALNT3.2591
36.781	0.759	CENPI.2491
-35.215	-0.634	FGD3.89846

Showing 1 to 10 of 59 entries   Previous 1 2 3 4 5 6 Next

List of all significant genes, in order of decreasing importance score are reported

Tab 4: Super PC Binary Outcome


GAC: Gene Associations with Clinical
Read Me
Super PC Time to Event Outcome
Super PC Continuous Outcome
Super PC Binary Outcome
Forest Plot
Tutorial
About Us ▾



Input file

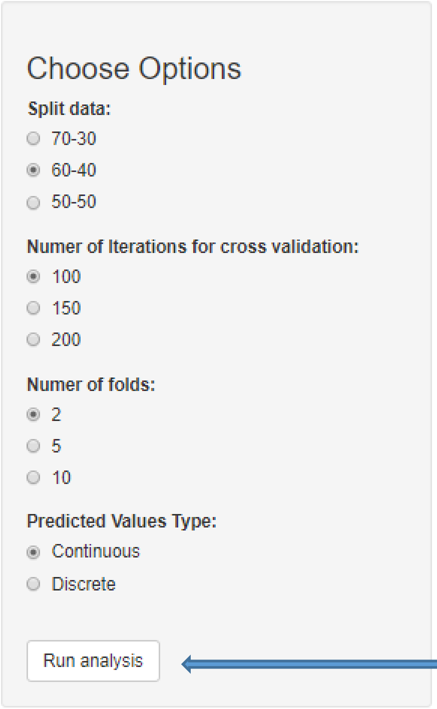
Select an example ds or upload your own with 'Load my own'

Example ds File ▾

Step 1:

Select example ‘expression with binary outcome indicator’ dataset or upload your own.

To view example data and format, use download button to view .csv file.



Choose Options

Split data:

○ 70-30
● 60-40
○ 50-50

Numer of Iterations for cross validation:

● 100
○ 150
○ 200

Numer of folds:

● 2
○ 5
○ 10

Predicted Values Type:

● Continuous
○ Discrete

Step 2:

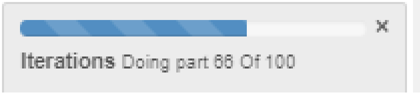
Choose data split into training and validation. Default uses 60-40 i.e. data is split into 40 % for training and remaining 60% for validation. You can also choose the number of interactions. For example purposes a smaller number 100 is used, but greater number of iterations are preferred.

Also, select number of folds for cross validation. Default is 2.

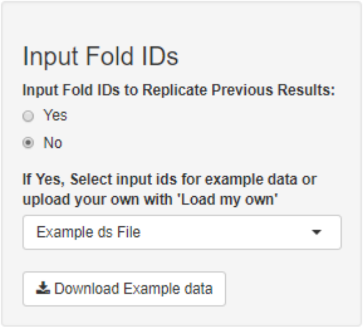
Additionally, choose to display predicted values as boxplots (for continuous predictors with t-test) or bar plot (for discrete groups cut off at median with chisq test).

Step 3:

Hit button to run analysis after each change in option. A progress indicator may appear to the right corner of the page.



Iterations Doing part 66 Of 100



Input Fold IDs

Input Fold IDs to Replicate Previous Results:

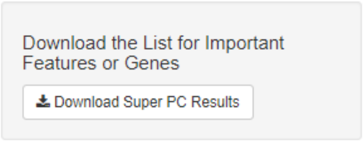
○ Yes
● No

If Yes, Select input ids for example data or upload your own with 'Load my own'

Example ds File ▾

Step 4 (Optional):

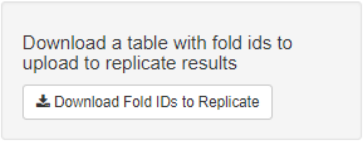
The user can replicate previously generated results by uploading fold-ids (download available below). When running analysis for first time, leave option to ‘No’.



Download the List for Important Features or Genes

Step 5 (Optional):

The user can download the tabular results in .csv file.

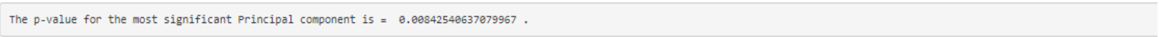


Download a table with fold ids to upload to replicate results

Step 6 (Optional):

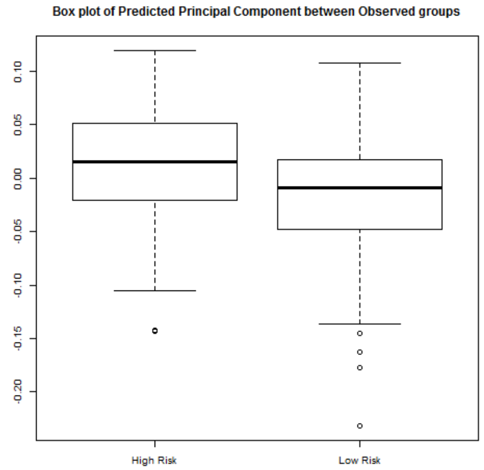
The user can download the fold-ids in csv format to replicate results at a later time. These can be uploaded in the input fold id section.

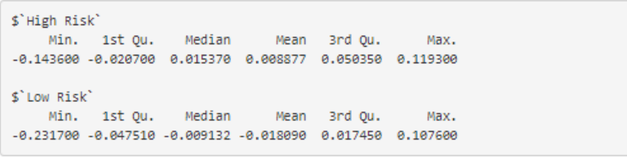
Association between Predicted Principal Component with Observed Outcome



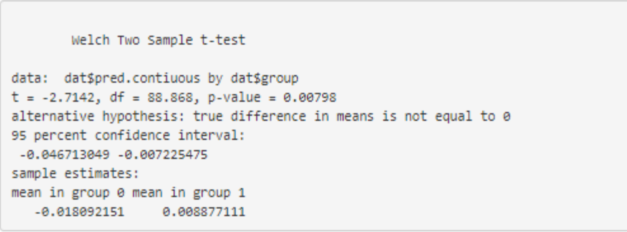
```
The p-value for the most significant Principal component is =  0.00842540637079967 .
```

Plots and Statistical Results for Predicted Principal Component between Observed Outcome





```
$`High Risk`
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.143600 -0.020700  0.015370  0.008877  0.050350  0.119300

$`Low Risk`
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.231700 -0.047510 -0.009132 -0.018090  0.017450  0.107600
```

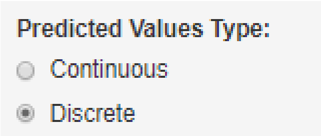


```

	Welch Two Sample t-test

data:  dat$pred.contiuous by dat$group
t = -2.7142, df = 88.868, p-value = 0.00798
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.046713049 -0.007225475
sample estimates:
mean in group 0 mean in group 1
   -0.018092151     0.008877111
```

Main panel shows the results from the super PC analysis.

The p-value from 1st supervised PC is reported.

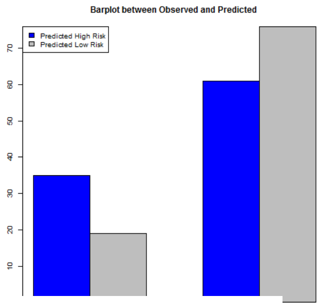
Scatter plot showing correlation of continuous predictor with outcome using Pearson correlation is reported. Discrete predictors are created by cutting the predictor at its median to form two groups and reported as boxplots.

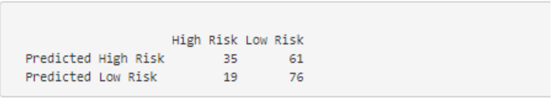


Predicted Values Type:

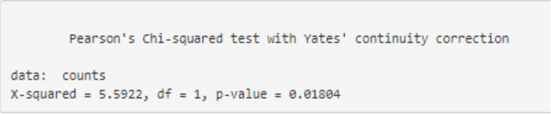
○ Continuous

● Discrete



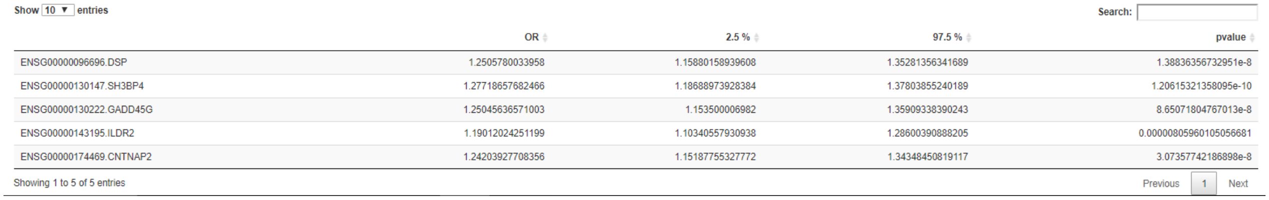


```
                    High Risk Low Risk
Predicted High Risk        35       61
Predicted Low Risk         19       76
```



```

	Pearson's Chi-squared test with Yates' continuity correction

data:  counts
X-squared = 5.5922, df = 1, p-value = 0.01804
```

Univariate logistic regresssion of genes associated with outcome to generate Principal components



Show 10 ▼ entries

Search:

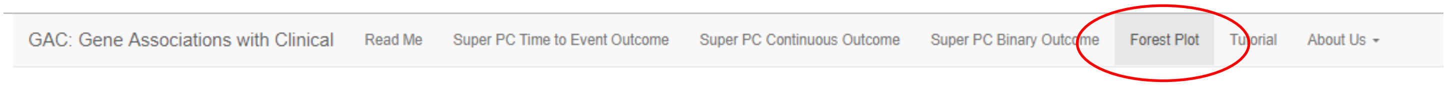
	OR	2.5 %	97.5 %	pvalue
ENSG00000096696.DSP	1.2505780033958	1.15880158939608	1.35281356341689	1.38836356732951e-8
ENSG00000130147.SH3BP4	1.27718657682466	1.18688973928384	1.37803855240189	1.20615321358095e-10
ENSG00000130222.GADD45G	1.25045636571003	1.153500006982	1.35909338390243	8.65071804767013e-8
ENSG00000143195.ILDR2	1.19012024251199	1.10340557930938	1.28600390888205	0.00000805960105056681
ENSG00000174469.CNTNAP2	1.24203927708356	1.15187755327772	1.34348450819117	3.07357742186898e-8

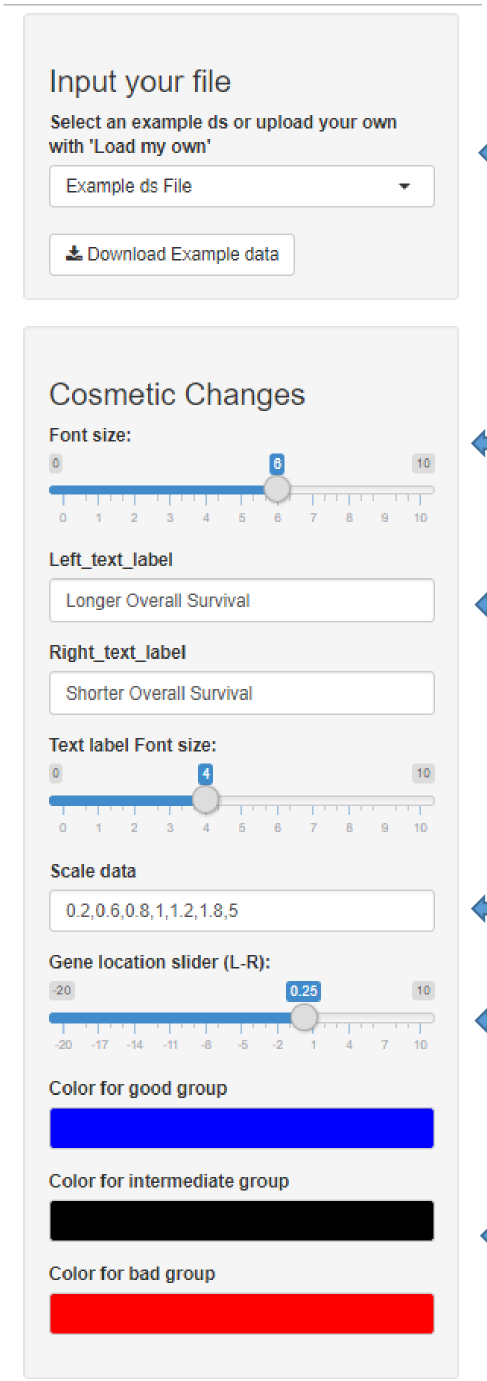
Showing 1 to 5 of 5 entries

Previous 1 Next

List of all significant genes, in order of decreasing importance score are reported

Tab 4: Forest Plot


GAC: Gene Associations with Clinical
Read Me
Super PC Time to Event Outcome
Super PC Continuous Outcome
Super PC Binary Outcome
Forest Plot
Tutorial
About Us


Input your file

Select an example ds or upload your own with 'Load my own'

Example ds File

Cosmetic Changes

Font size: 6 (0–10)

Left_text_label: Longer Overall Survival

Right_text_label: Shorter Overall Survival

Text label Font size: 4 (0–10)

Scale data: 0.2,0.6,0.8,1,1.2,1.8,5

Gene location slider (L-R): 0.25 (-20–10)

Color for good group: 🟦

Color for intermediate group: ⬛

Color for bad group: 🟥

Step 1:

Select example dataset or upload your own.

To view example data and format, use download button to view .csv file.

Step 2:

Change font size for the IDs (genes) displayed to the left of forest plot.

Step 3:

Change label for text to be displayed above the horizontal line to the left and right of the HR= 1 dotted line.

A slider bar to change font size for these labels is also available.

Step 4:

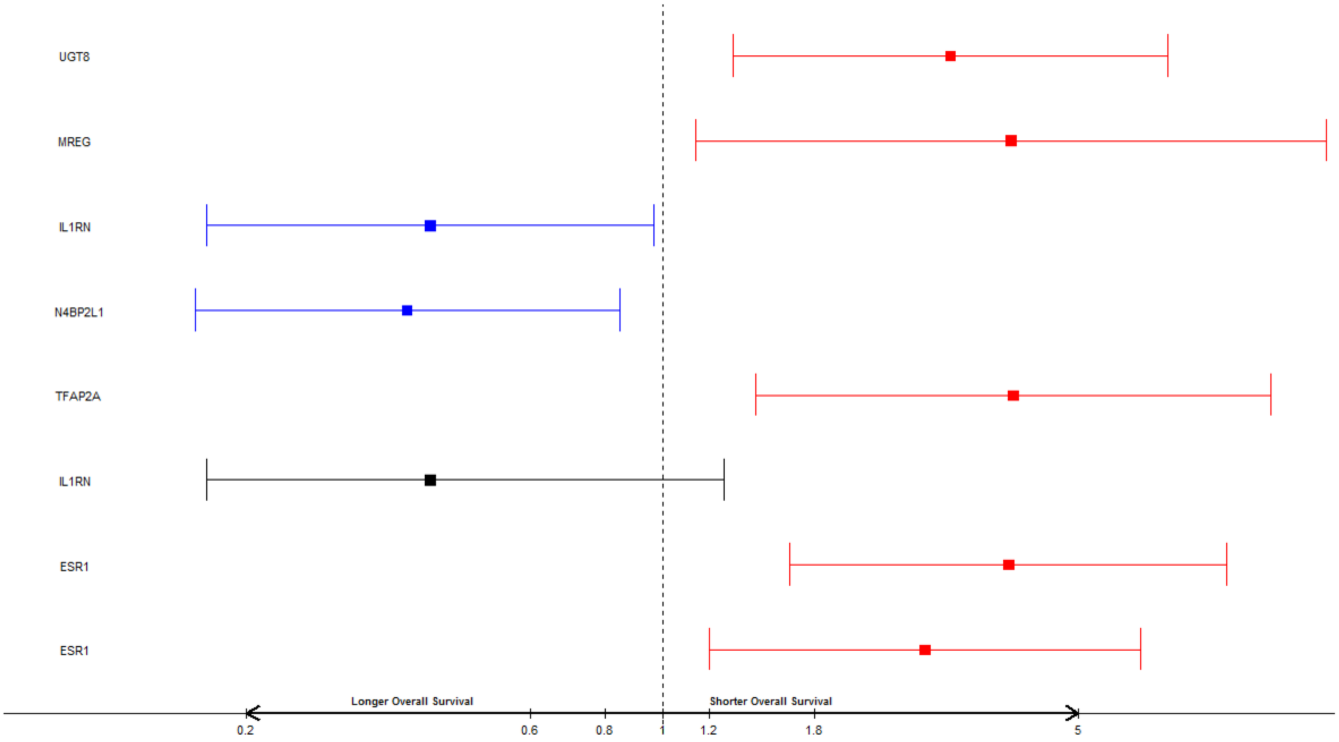
Choose x ticks to display below the horizontal line depending on dataset.

Step 7:

Move IDs (genes) slightly to right or left to avoid overlap with the forest plot.

Step 8:

Change color of outcomes. Genes with HR and their 95% CI below 1 are coded blue (good group), overlapping 1 are coded black (the intermediate group) and those over 1 are coded red (the bad group).



Reporting results from super PC analysis through forest plot. Genes with HR < 1 are coded in blue, those overlapping 1 are coded black and those greater than 1 are coded red.