Jun 06, 2019 Version 2

Machine learning approach yields epigenetic biomarkers of food allergy: A novel 13-gene signature to diagnose clinical reactivity V.2

PLOS One

DOI
dx.doi.org/10.17504/protocols.io.x7pfrmn

Ayush Alag[1]

[1]The Harker School

Ayush Alag

DOI: dx.doi.org/10.17504/protocols.io.x7pfrmn

External link: https://doi.org/10.1371/journal.pone.0218253

Protocol Citation: Ayush Alag 2019. Machine learning approach yields epigenetic biomarkers of food allergy: A novel 13-gene signature to diagnose clinical reactivity. **protocols.io https://dx.doi.org/10.17504/protocols.io.x7pfrmn**

Manuscript citation:

Protocol status: Working

Created: February 15, 2019

Last Modified: June 06, 2019

Protocol Integer ID: 20431

Keywords: Machine Learning, Food Allergy, DNA Methylation, Gene Signature

Abstract

Current laboratory tests have a less than 50% accuracy in distinguishing between people who have food allergies (FA) and those who are merely sensitized to foods, resulting in the use of expensive and potentially dangerous Oral Food Challenges. Our study presents a purely-computational machine learning approach, conducted using DNA Methylation (DNAm) data, to accurately diagnose food allergies and find genes that are strong biomarkers of the disease.

We built two deep learning classifiers with twelve CpG-input features each that achieved perfect accuracy and an AUROC of 1 on the completely hidden cross-validation cohort. In addition, 24 additional classifiers were created that each had an average cross-validation accuracy of 98.35%. These 26 classifiers yielded a total of 18 unique CpGs, which mapped to 13 genes that are strong epigenetic biomarkers of FA.

Biological enrichment on the 13-gene signature yielded new insights. Notably, our FA-discriminating genes were strongly associated with the immune system, which helps validate our findings. Seven of the 13 genes overlapped with previous food-allergy and DNAm studies.

Previous studies have also created a perfect classifier for this dataset, but they used a 96-CpG input feature set built on both data-driven and a priori biological insights. Our study is an improvement on previous work because it maintains a perfect classification accuracy using only 18 highly discriminating CpGs (0.005% of the total available features). In machine learning, simpler models, as used in our study, are preferred over more complex ones (all other things being equal).

In addition, our completely data-driven approach eliminates the need for \textit{a priori} information and allows for generalizability to DNAm classification problems in other disease areas, which may result in novel gene associations or accurate diagnostic tests for those diseases.

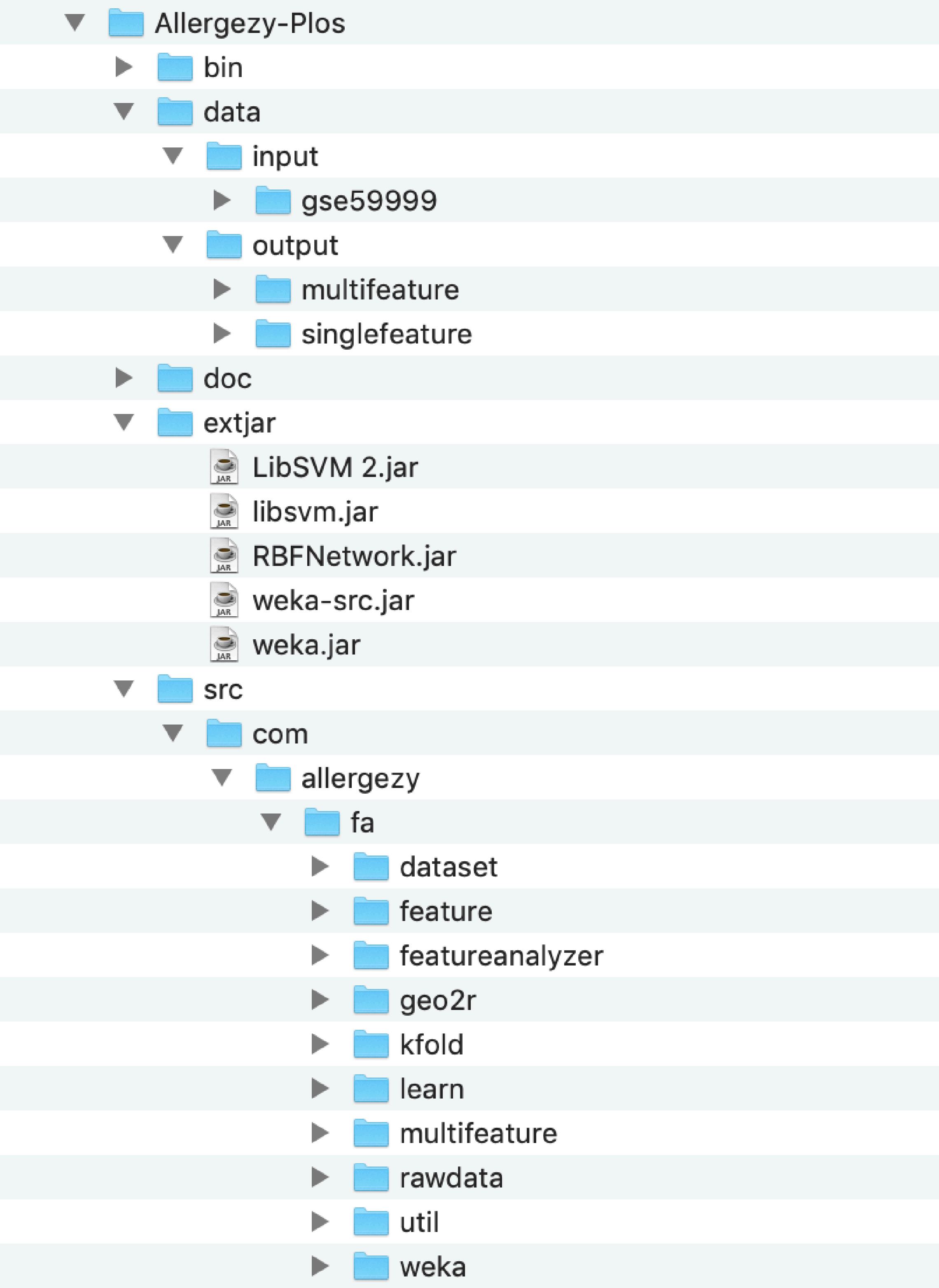
Guidelines

The DNA Methylation data analysis has been done using the Java programming language. To follow the steps it is best to create an Eclipse project.

Before start

1. Download the Eclipse IDE (https://www.eclipse.org/downloads/)

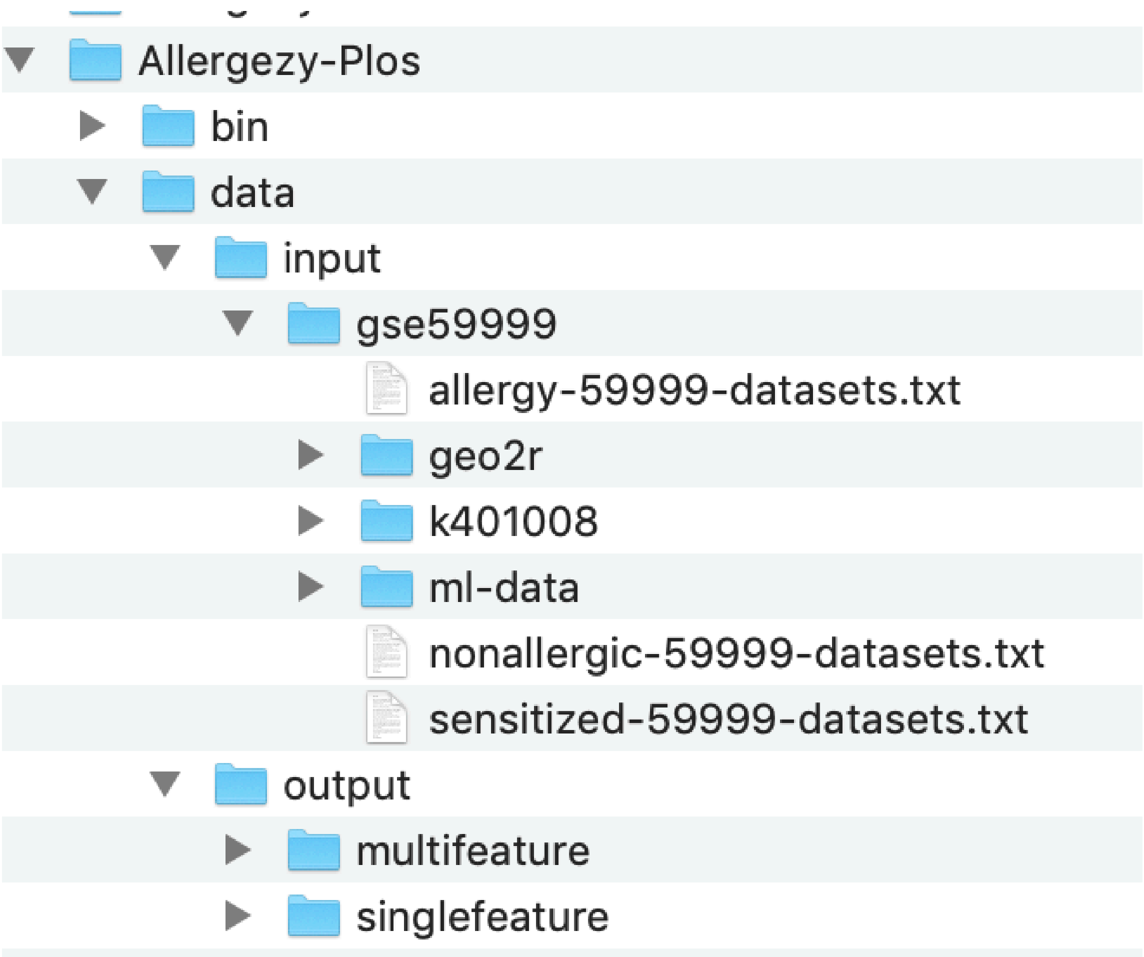
1 Create an Eclipse project. Download weka jar files from https://www.cs.waikato.ac.nz/ml/weka/downloading.html . The Eclipse project should have the src files with the Java code downloaded in the previous step. Create a directory extjar and put the two reference jar files from Weka (weka-src.jar and weka.jar). Download RBFNetwork.jar from https://jar-download.com/artifact-search/RBFNetwork and add it to the extjar directory. Download libsvm.jar from https://jar-download.com/download-handling.php. Download the data.jar file and place it within the main project. This directory should have the input and output folders with many sub folders. The doc folder is optional and creates the generated JavaDoc for the project. JavaDoc can be generated by going to Eclipse → Project → Generate JavaDoc.



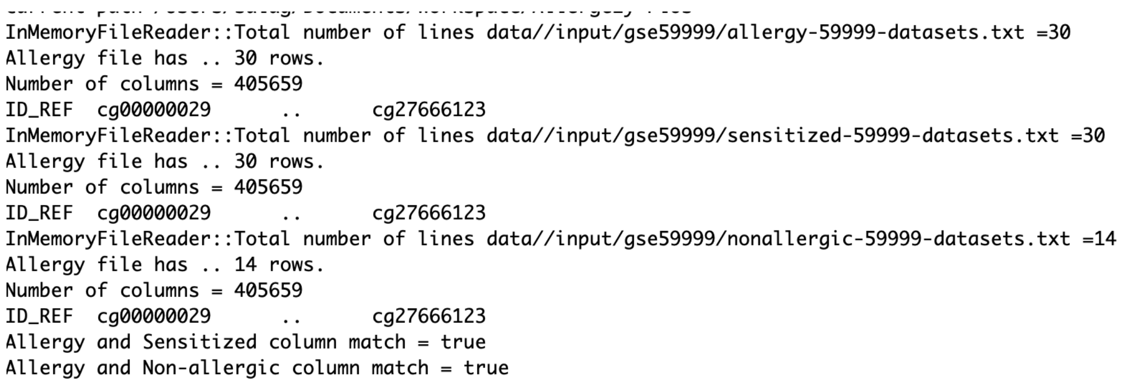
```
▼ Allergezy-Plos
    ▶ bin
    ▼ data
        ▼ input
            ▶ gse59999
        ▼ output
            ▶ multifeature
            ▶ singlefeature
    ▶ doc
    ▼ extjar
          LibSVM 2.jar
          libsvm.jar
          RBFNetwork.jar
          weka-src.jar
          weka.jar
    ▼ src
        ▼ com
            ▼ allergezy
                ▼ fa
                    ▶ dataset
                    ▶ feature
                    ▶ featureanalyzer
                    ▶ geo2r
                    ▶ kfold
                    ▶ learn
                    ▶ multifeature
                    ▶ rawdata
                    ▶ util
                    ▶ weka
```

2 Next, we will validate that the CpG data that has already been separated into Food Allergy, Sensitized, and Non-Allergic samples have been downloaded correctly. Expand the data/input/gse59999 folder and verify that there are three files:

- allergy-59999-datasets.txt 215 MB contains 30 rows by 405,659 columns
- nonallergic-59999-datasets.txt 98.9 MB contains 14 rows by 405,659 columns
- sensitized-59999-datasets.txt 215 MB contains 30 rows by 405,659 columns


Allergezy-Plos
bin
data
input
gse59999
allergy-59999-datasets.txt
geo2r
k401008
ml-data
nonallergic-59999-datasets.txt
sensitized-59999-datasets.txt
output
multifeature
singlefeature

From within Eclipse, run the main program for com.allergezy.fa.util.FileUtil. You should see the output shown below.
Note the number of rows and columns and the validation check.



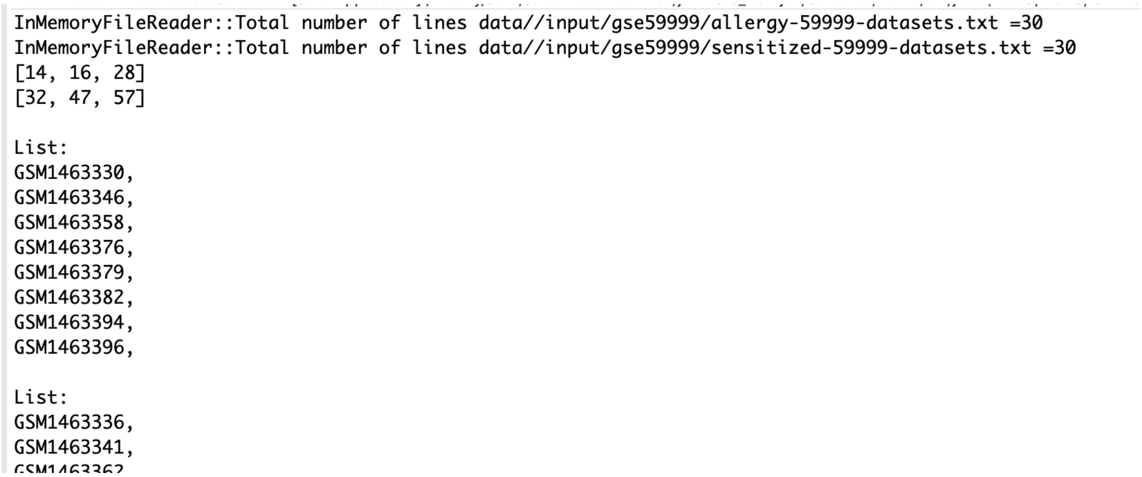
```
InMemoryFileReader::Total number of lines data//input/gse59999/allergy-59999-datasets.txt =30
Allergy file has .. 30 rows.
Number of columns = 405659
ID_REF  cg00000029      ..      cg27666123
InMemoryFileReader::Total number of lines data//input/gse59999/sensitized-59999-datasets.txt =30
Allergy file has .. 30 rows.
Number of columns = 405659
ID_REF  cg00000029      ..      cg27666123
InMemoryFileReader::Total number of lines data//input/gse59999/nonallergic-59999-datasets.txt =14
Allergy file has .. 14 rows.
Number of columns = 405659
ID_REF  cg00000029      ..      cg27666123
Allergy and Sensitized column match = true
Allergy and Non-allergic column match = true
```

The research was conducted using a dataset found in the Gene Expression Omnibus (GEO)under accession id GSE59999. The 71 patient samples in this dataset consisted of 29 patients with FA (tested positive on OFCs), 29 patients who were sensitized but not food-allergic, and 13 patients who were neither sensitized nor allergic. Each sample is associated with a DNAm profile that was taken from mononuclear blood cells and consisted of normalized Beta values ranging from 0 (completely unmethylated) to 1 (fully methylated) at 405,658 CpG islands distributed across the genome.

3 In this step, we will create 8-fold test/train/hidden samples.

The 58 samples are randomly split into three datasets: 40 samples for training, 10 samples for testing, and 8 completely hidden samples for cross-validation. Half of the samples in each of the three datasets were children with food allergies and the other half were those that were food-sensitized but not food-allergic. To avoid potential bias, we created eight random splits, shuffling the samples across the three datasets each time such that each of the 58 samples was in the hidden dataset at least once.

The logic is implemented in com.allergezy.fa.kfold.KFoldDatasetGenerator. Run the main method in this class to see an output similar to below.

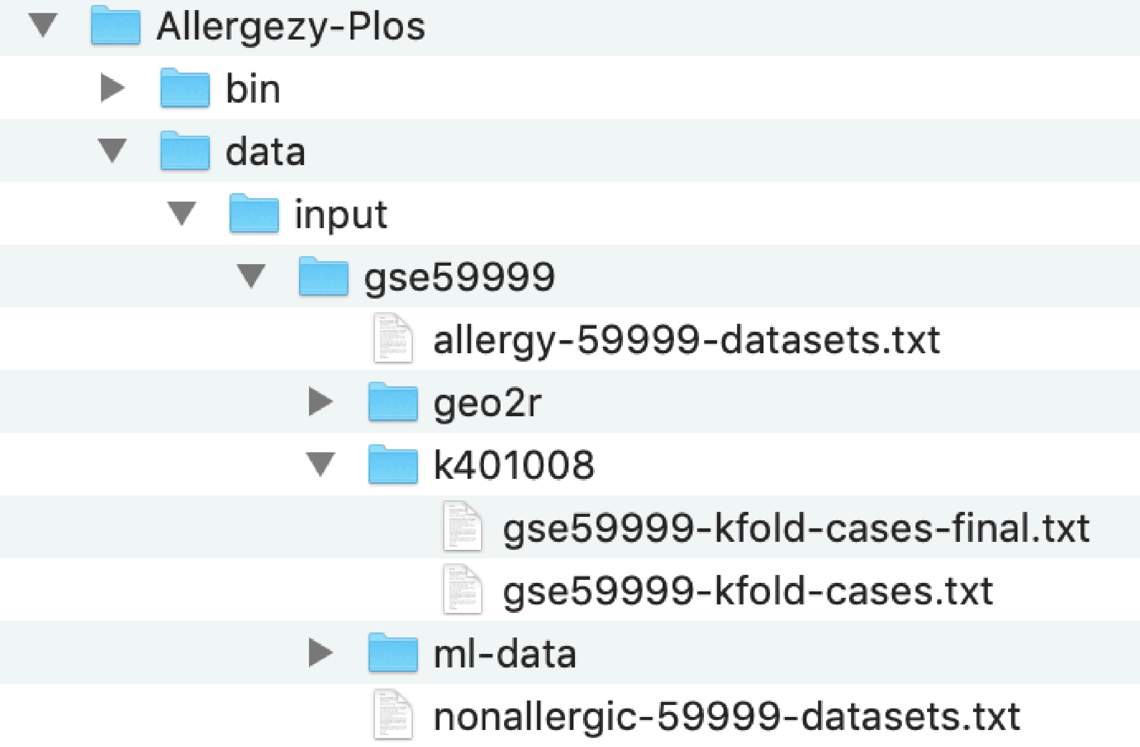


```
InMemoryFileReader::Total number of lines data//input/gse59999/allergy-59999-datasets.txt =30
InMemoryFileReader::Total number of lines data//input/gse59999/sensitized-59999-datasets.txt =30
[14, 16, 28]
[32, 47, 57]

List:
GSM1463330,
GSM1463346,
GSM1463358,
GSM1463376,
GSM1463379,
GSM1463382,
GSM1463394,
GSM1463396,

List:
GSM1463336,
GSM1463341,
GSM1463362
```

Output from running com.allergezy.fa.kfold.KFoldDatasetGenerator.

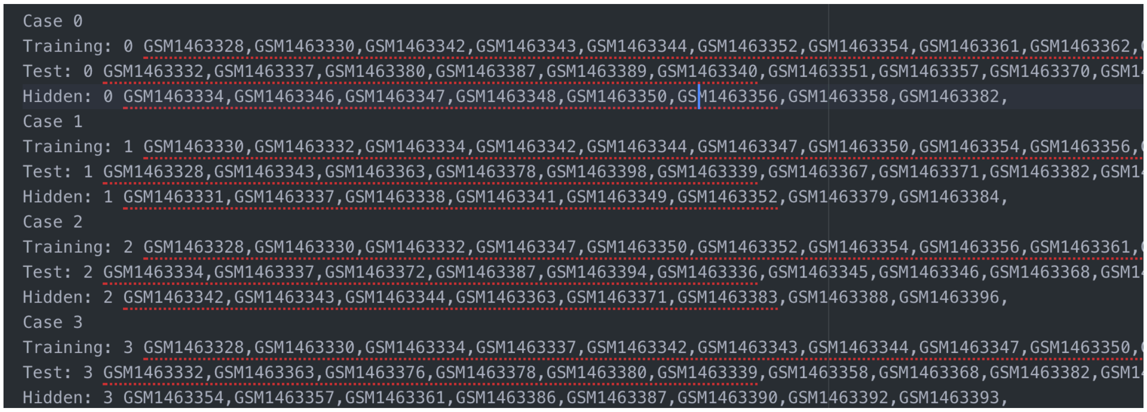
This program produces an output file data/input/gse59999/k401008/gse59999-kfold-cases.txt as shown below.

The file data/input/gse59999/k401008/gse59999-kfold-cases-final.txt is the final version of this 8-fold split that is used in the rest of the study.



```
Allergezy-Plos
    bin
    data
        input
            gse59999
                allergy-59999-datasets.txt
                geo2r
                k401008
                    gse59999-kfold-cases-final.txt
                    gse59999-kfold-cases.txt
                ml-data
                nonallergic-59999-datasets.txt
```

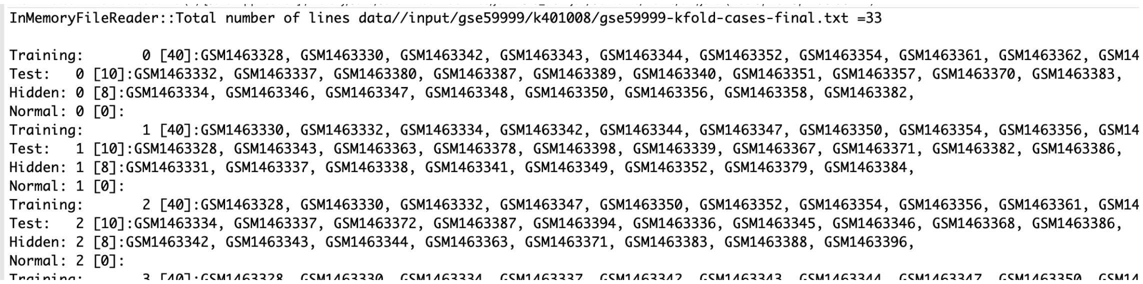
Below is the content of the data/input/gse59999/k401008/gse59999-kfold-cases-final.txt file.



```
Case 0
Training: 0 GSM1463328,GSM1463330,GSM1463342,GSM1463343,GSM1463344,GSM1463352,GSM1463354,GSM1463361,GSM1463362,
Test: 0 GSM1463332,GSM1463337,GSM1463380,GSM1463387,GSM1463389,GSM1463340,GSM1463351,GSM1463357,GSM1463370,GSM1
Hidden: 0 GSM1463334,GSM1463346,GSM1463347,GSM1463348,GSM1463350,GSM1463356,GSM1463358,GSM1463382,
Case 1
Training: 1 GSM1463330,GSM1463332,GSM1463334,GSM1463342,GSM1463344,GSM1463347,GSM1463350,GSM1463354,GSM1463356,
Test: 1 GSM1463328,GSM1463343,GSM1463363,GSM1463378,GSM1463398,GSM1463339,GSM1463367,GSM1463371,GSM1463382,GSM1
Hidden: 1 GSM1463331,GSM1463337,GSM1463338,GSM1463341,GSM1463349,GSM1463352,GSM1463379,GSM1463384,
Case 2
Training: 2 GSM1463328,GSM1463330,GSM1463332,GSM1463347,GSM1463350,GSM1463352,GSM1463354,GSM1463356,GSM1463361,
Test: 2 GSM1463334,GSM1463337,GSM1463372,GSM1463387,GSM1463394,GSM1463336,GSM1463345,GSM1463346,GSM1463368,GSM1
Hidden: 2 GSM1463342,GSM1463343,GSM1463344,GSM1463363,GSM1463371,GSM1463383,GSM1463388,GSM1463396,
Case 3
Training: 3 GSM1463328,GSM1463330,GSM1463334,GSM1463337,GSM1463342,GSM1463343,GSM1463344,GSM1463347,GSM1463350,
Test: 3 GSM1463332,GSM1463363,GSM1463376,GSM1463378,GSM1463380,GSM1463339,GSM1463358,GSM1463368,GSM1463382,GSM1
Hidden: 3 GSM1463354,GSM1463357,GSM1463361,GSM1463386,GSM1463387,GSM1463390,GSM1463392,GSM1463393,
```

Each sample appears at least once in the hidden set. In addition, for every case there are 40 training samples. There is an equal number of food-allergic and food-sensitized samples in each dataset and in each case.

To verify that this file can be accessed, run the main method for com.allergezy.fa.kfold.KFoldDatasetInfo. This file follows the Singleton design pattern and provides information on which samples are in the training, test, and hidden datasets for each of the eight cases. You should see the following output.



```
InMemoryFileReader::Total number of lines data//input/gse59999/k401008/gse59999-kfold-cases-final.txt =33

Training:       0 [40]:GSM1463328, GSM1463330, GSM1463342, GSM1463343, GSM1463344, GSM1463352, GSM1463354, GSM1463361, GSM1463362, GSM14
Test:   0 [10]:GSM1463332, GSM1463337, GSM1463380, GSM1463387, GSM1463389, GSM1463340, GSM1463351, GSM1463357, GSM1463370, GSM1463383,
Hidden: 0 [8]:GSM1463334, GSM1463346, GSM1463347, GSM1463348, GSM1463350, GSM1463356, GSM1463358, GSM1463382,
Normal: 0 [0]:
Training:       1 [40]:GSM1463330, GSM1463332, GSM1463334, GSM1463342, GSM1463344, GSM1463347, GSM1463350, GSM1463354, GSM1463356, GSM14
Test:   1 [10]:GSM1463328, GSM1463343, GSM1463363, GSM1463378, GSM1463398, GSM1463339, GSM1463367, GSM1463371, GSM1463382, GSM1463386,
Hidden: 1 [8]:GSM1463331, GSM1463337, GSM1463338, GSM1463341, GSM1463349, GSM1463352, GSM1463379, GSM1463384,
Normal: 1 [0]:
Training:       2 [40]:GSM1463328, GSM1463330, GSM1463332, GSM1463347, GSM1463350, GSM1463352, GSM1463354, GSM1463356, GSM1463361, GSM14
Test:   2 [10]:GSM1463334, GSM1463337, GSM1463372, GSM1463387, GSM1463394, GSM1463336, GSM1463345, GSM1463346, GSM1463368, GSM1463386,
Hidden: 2 [8]:GSM1463342, GSM1463343, GSM1463344, GSM1463363, GSM1463371, GSM1463383, GSM1463388, GSM1463396,
Normal: 2 [0]:
Training:       3 [40]:GSM1463328, GSM1463330, GSM1463334, GSM1463337, GSM1463342, GSM1463343, GSM1463344, GSM1463347, GSM1463350, GSM14
```

The distribution of the samples can be seen in the following table.

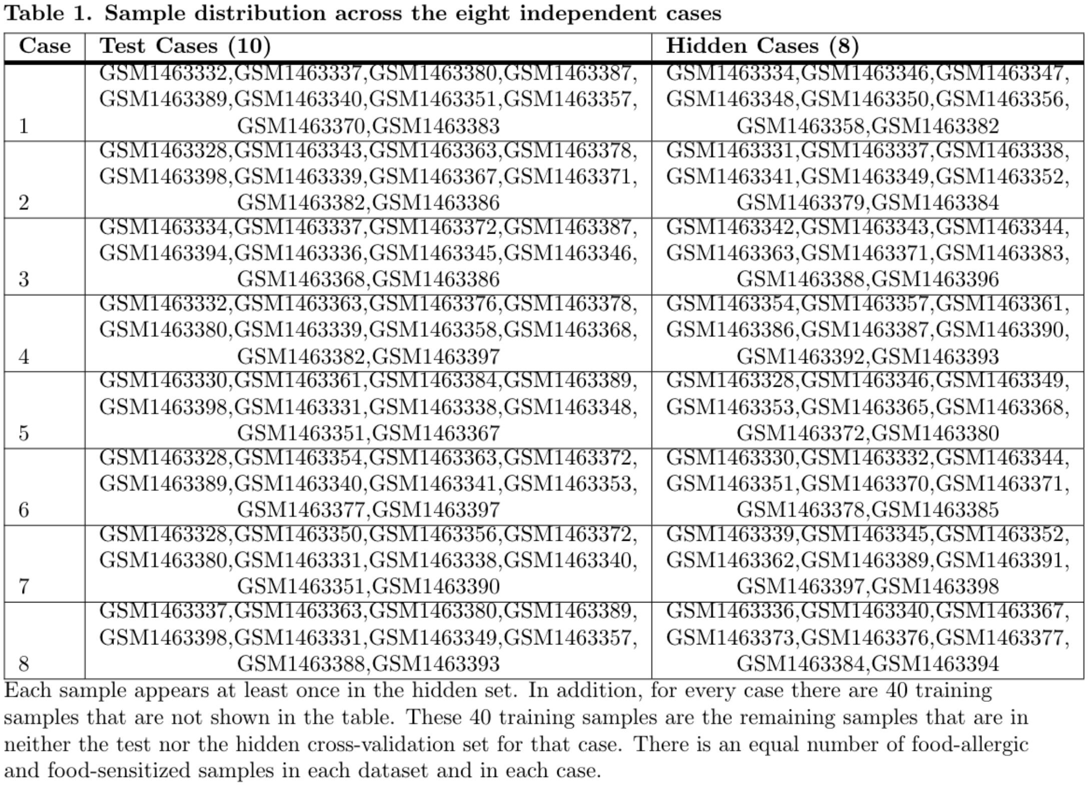


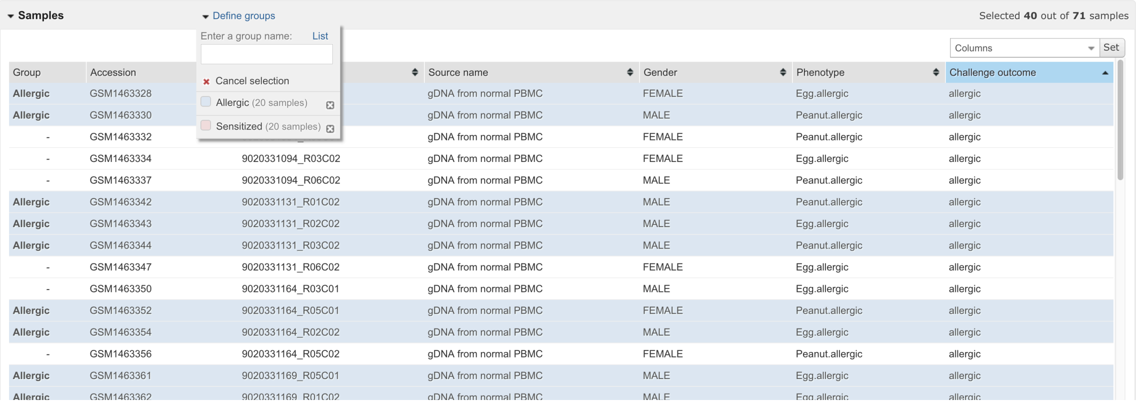
Table 1. Sample distribution across the eight independent cases

Case	Test Cases (10)	Hidden Cases (8)
1	GSM1463332,GSM1463337,GSM1463380,GSM1463387, GSM1463389,GSM1463340,GSM1463351,GSM1463357, GSM1463370,GSM1463383	GSM1463334,GSM1463346,GSM1463347, GSM1463348,GSM1463350,GSM1463356, GSM1463358,GSM1463382
2	GSM1463328,GSM1463343,GSM1463363,GSM1463378, GSM1463398,GSM1463339,GSM1463367,GSM1463371, GSM1463382,GSM1463386	GSM1463331,GSM1463337,GSM1463338, GSM1463341,GSM1463349,GSM1463352, GSM1463379,GSM1463384
3	GSM1463334,GSM1463337,GSM1463372,GSM1463387, GSM1463394,GSM1463336,GSM1463345,GSM1463346, GSM1463368,GSM1463386	GSM1463342,GSM1463343,GSM1463344, GSM1463363,GSM1463371,GSM1463383, GSM1463388,GSM1463396
4	GSM1463332,GSM1463363,GSM1463376,GSM1463378, GSM1463380,GSM1463339,GSM1463358,GSM1463368, GSM1463382,GSM1463397	GSM1463354,GSM1463357,GSM1463361, GSM1463386,GSM1463387,GSM1463390, GSM1463392,GSM1463393
5	GSM1463330,GSM1463361,GSM1463384,GSM1463389, GSM1463398,GSM1463331,GSM1463338,GSM1463348, GSM1463351,GSM1463367	GSM1463328,GSM1463346,GSM1463349, GSM1463353,GSM1463365,GSM1463368, GSM1463372,GSM1463380
6	GSM1463328,GSM1463354,GSM1463363,GSM1463372, GSM1463389,GSM1463340,GSM1463341,GSM1463353, GSM1463377,GSM1463397	GSM1463330,GSM1463332,GSM1463344, GSM1463351,GSM1463370,GSM1463371, GSM1463378,GSM1463385
7	GSM1463328,GSM1463350,GSM1463356,GSM1463372, GSM1463380,GSM1463331,GSM1463338,GSM1463340, GSM1463351,GSM1463390	GSM1463339,GSM1463345,GSM1463352, GSM1463362,GSM1463389,GSM1463391, GSM1463397,GSM1463398
8	GSM1463337,GSM1463363,GSM1463380,GSM1463389, GSM1463398,GSM1463331,GSM1463349,GSM1463357, GSM1463388,GSM1463393	GSM1463336,GSM1463340,GSM1463367, GSM1463373,GSM1463376,GSM1463377, GSM1463384,GSM1463394

Each sample appears at least once in the hidden set. In addition, for every case there are 40 training samples that are not shown in the table. These 40 training samples are the remaining samples that are in neither the test nor the hidden cross-validation set for that case. There is an equal number of food-allergic and food-sensitized samples in each dataset and in each case.

4 DNAm datasets are characterized as having a small number of samples but a very high number of feature dimensions (HDLSS). To prevent overfitting and increase generalization, it is important to condense the feature list relative to the number of samples available. Computationally, it is very expensive to evaluate the more than 400K CpG features individually. Therefore, in order to limit the evaluation size and begin with a list of potentially highly-relevant CpG points, we used the NCBI **GEO2R tool** to obtain a prioritized list of CpG features differentially expressed across the two groups. For each of the eight independent cases, we split the forty training samples into two cohorts, one with allergic patients and the other with sensitized patients.The GEO2R tool was used to derive eight lists of 100 CpGs each, one for each of the eight cases.

The screenshot below shows the use of the GEO2R tool to create two cohorts of 20 samples each for the first case.
Clicking on



▾ Samples ▾ Define groups Selected **40** out of **71** samples

Enter a group name: List

✖ Cancel selection

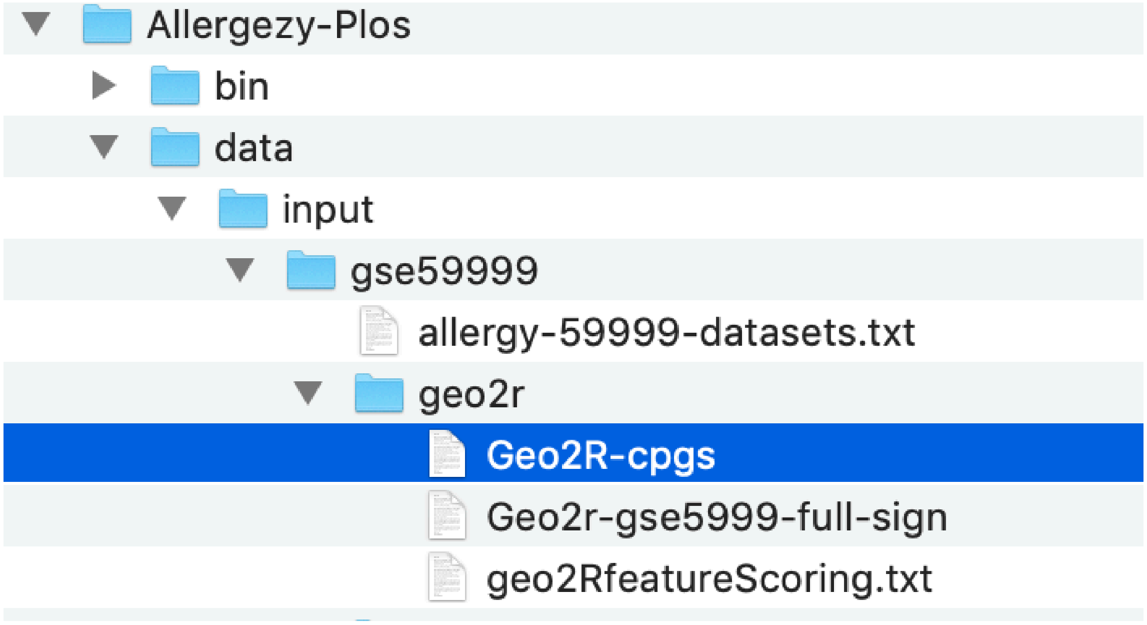
□ Allergic (20 samples)

□ Sensitized (20 samples)

Columns ▾ Set

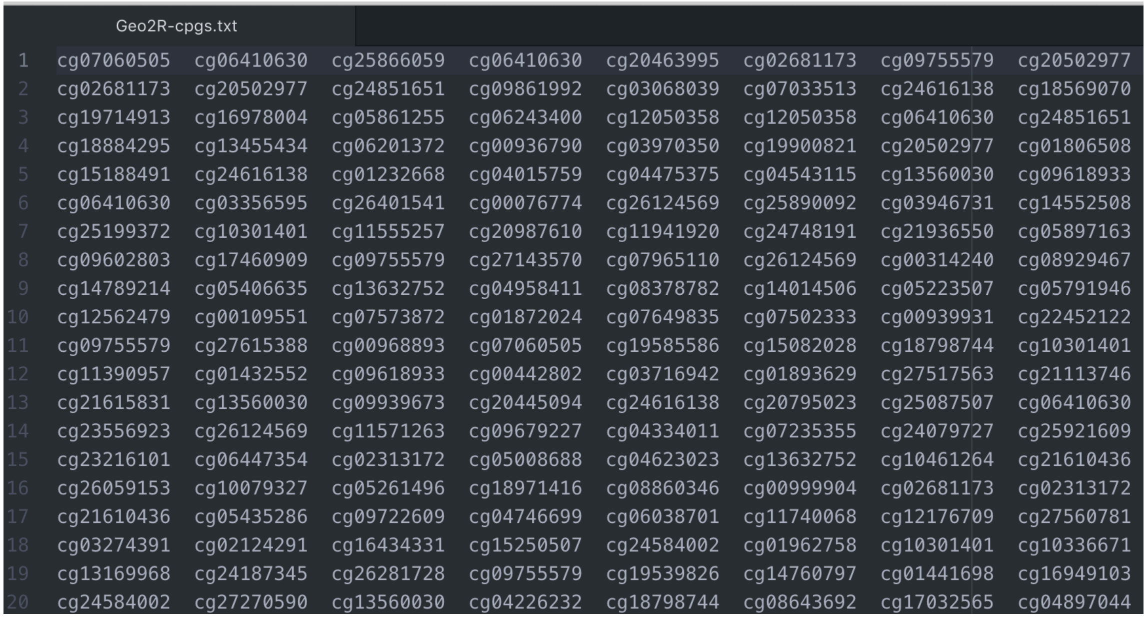
Group	Accession		Source name	Gender	Phenotype	Challenge outcome
Allergic	GSM1463328		gDNA from normal PBMC	FEMALE	Egg.allergic	allergic
Allergic	GSM1463330		gDNA from normal PBMC	MALE	Peanut.allergic	allergic
-	GSM1463332	[illegible]	gDNA from normal PBMC	FEMALE	Peanut.allergic	allergic
-	GSM1463334	9020331094_R03C02	gDNA from normal PBMC	FEMALE	Egg.allergic	allergic
-	GSM1463337	9020331094_R06C02	gDNA from normal PBMC	MALE	Peanut.allergic	allergic
Allergic	GSM1463342	9020331131_R01C02	gDNA from normal PBMC	MALE	Peanut.allergic	allergic
Allergic	GSM1463343	9020331131_R02C02	gDNA from normal PBMC	MALE	Egg.allergic	allergic
Allergic	GSM1463344	9020331131_R03C02	gDNA from normal PBMC	MALE	Peanut.allergic	allergic
-	GSM1463347	9020331131_R06C02	gDNA from normal PBMC	FEMALE	Egg.allergic	allergic
-	GSM1463350	9020331164_R03C01	gDNA from normal PBMC	MALE	Egg.allergic	allergic
Allergic	GSM1463352	9020331164_R05C01	gDNA from normal PBMC	FEMALE	Peanut.allergic	allergic
Allergic	GSM1463354	9020331164_R02C02	gDNA from normal PBMC	MALE	Egg.allergic	allergic
-	GSM1463356	9020331164_R05C02	gDNA from normal PBMC	FEMALE	Peanut.allergic	allergic
Allergic	GSM1463361	9020331169_R05C01	gDNA from normal PBMC	MALE	Egg.allergic	allergic
Allergic	GSM1463362	9020331169_R01C02	gDNA from normal PBMC	MALE	Egg.allergic	allergic

The full signature for the eight cases is in data/input/gse59999/geo2r/Geo2R-cpgs.txt file.



```
Allergezy-Plos
    bin
    data
        input
            gse59999
                allergy-59999-datasets.txt
                geo2r
                    Geo2R-cpgs
                    Geo2r-gse5999-full-sign
                    geo2RfeatureScoring.txt
```

Below are the first twenty lines of the file data/input/gse59999/geo2r/Geo2R-cpgs.txt file. This will be used in the next few steps.

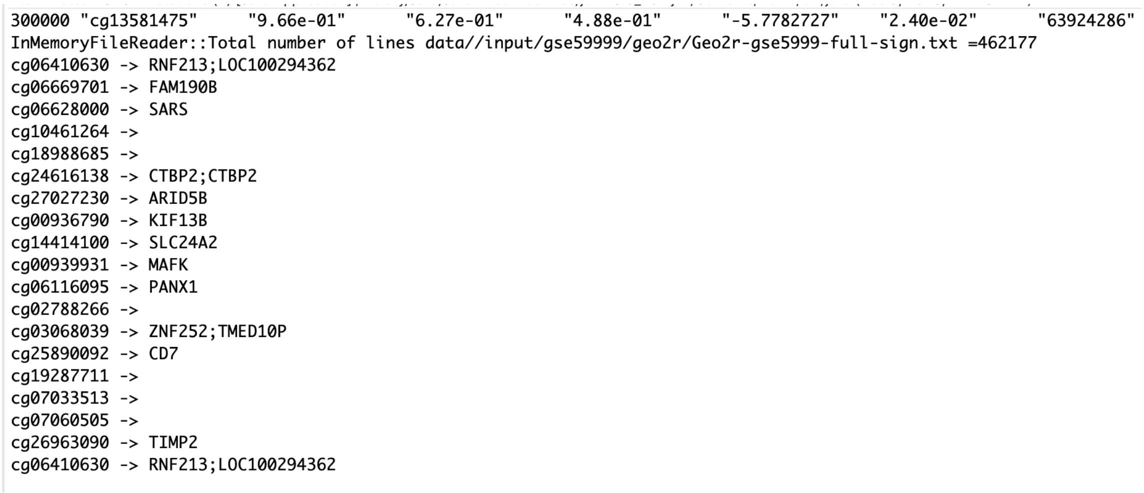


Geo2R-cpgs.txt

```
cg07060505  cg06410630  cg25866059  cg06410630  cg20463995  cg02681173  cg09755579  cg20502977
cg02681173  cg20502977  cg24851651  cg09861992  cg03068039  cg07033513  cg24616138  cg18569070
cg19714913  cg16978004  cg05861255  cg06243400  cg12050358  cg12050358  cg06410630  cg24851651
cg18884295  cg13455434  cg06201372  cg00936790  cg03970350  cg19900821  cg20502977  cg01806508
cg15188491  cg24616138  cg01232668  cg04015759  cg04475375  cg04543115  cg13560030  cg09618933
cg06410630  cg03356595  cg26401541  cg00076774  cg26124569  cg25890092  cg03946731  cg14552508
cg25199372  cg10301401  cg11555257  cg20987610  cg11941920  cg24748191  cg21936550  cg05897163
cg09602803  cg17460909  cg09755579  cg27143570  cg07965110  cg26124569  cg00314240  cg08929467
cg14789214  cg05406635  cg13632752  cg04958411  cg08378782  cg14014506  cg05223507  cg05791946
cg12562479  cg00109551  cg07573872  cg01872024  cg07649835  cg07502333  cg00939931  cg22452122
cg09755579  cg27615388  cg00968893  cg07060505  cg19585586  cg15082028  cg18798744  cg10301401
cg11390957  cg01432552  cg09618933  cg00442802  cg03716942  cg01893629  cg27517563  cg21113746
cg21615831  cg13560030  cg09939673  cg20445094  cg24616138  cg20795023  cg25087507  cg06410630
cg23556923  cg26124569  cg11571263  cg09679227  cg04334011  cg07235355  cg24079727  cg25921609
cg23216101  cg06447354  cg02313172  cg05008688  cg04623023  cg13632752  cg10461264  cg21610436
cg26059153  cg10079327  cg05261496  cg18971416  cg08860346  cg00999904  cg02681173  cg02313172
cg21610436  cg05435286  cg09722609  cg04746699  cg06038701  cg11740068  cg12176709  cg27560781
cg03274391  cg02124291  cg16434331  cg15250507  cg24584002  cg01962758  cg10301401  cg10336671
cg13169968  cg24187345  cg26281728  cg09755579  cg19539826  cg14760797  cg01441698  cg16949103
cg24584002  cg27270590  cg13560030  cg04226232  cg18798744  cg08643692  cg17032565  cg04897044
```

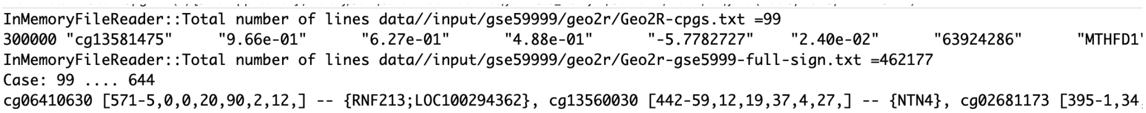
5 Next, the eight lists were combined together, which resulted in 636 unique CpGs, the count being less than 800 as some of the CpGs were repeated across the eight cases.

To map the CpGs to genes, we use the output from GEO2R contained in the file data/input/gse59999/geo2r/Geo2r-gse5999-full-sign.txt. The Singleton class com.allergezy.fa.geo2r.CPGAnnotations maps a CpG to genes using this file. Run the main method in the class com.allergezy.fa.geo2r.CPGAnnotations. You should see the output similar to the screen shot below.



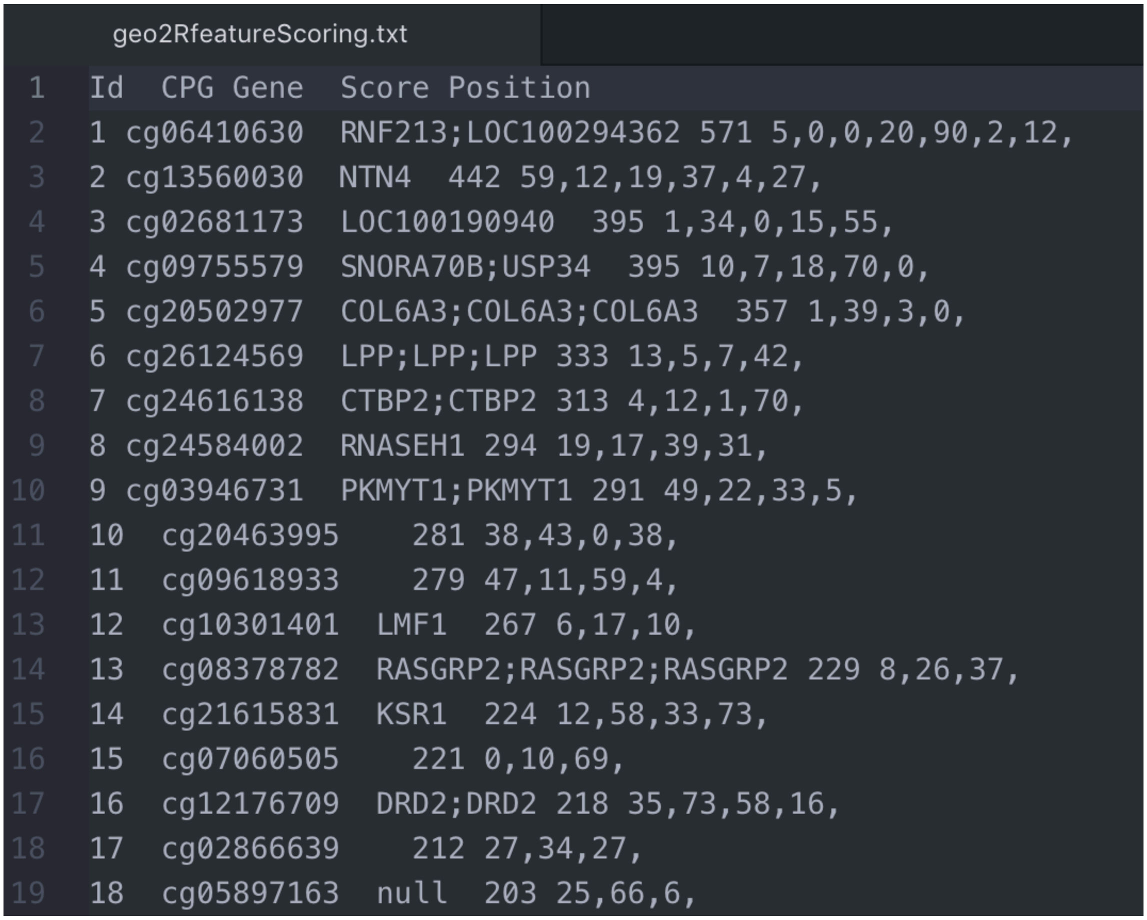
```
300000 "cg13581475"     "9.66e-01"      "6.27e-01"      "4.88e-01"      "-5.7782727"    "2.40e-02"      "63924286"
InMemoryFileReader::Total number of lines data//input/gse59999/geo2r/Geo2r-gse5999-full-sign.txt =462177
cg06410630 -> RNF213;LOC100294362
cg06669701 -> FAM190B
cg06628000 -> SARS
cg10461264 ->
cg18988685 ->
cg24616138 -> CTBP2;CTBP2
cg27027230 -> ARID5B
cg00936790 -> KIF13B
cg14414100 -> SLC24A2
cg00939931 -> MAFK
cg06116095 -> PANX1
cg02788266 ->
cg03068039 -> ZNF252;TMED10P
cg25890092 -> CD7
cg19287711 ->
cg07033513 ->
cg07060505 ->
cg26963090 -> TIMP2
cg06410630 -> RNF213;LOC100294362
```

The class com.allergezy.fa.geo2r.Geo2rCpgInfo computes the overlap between the CpGs from the GEO2R output. Running the main method produces the output shown below.



```
InMemoryFileReader::Total number of lines data//input/gse59999/geo2r/Geo2R-cpgs.txt =99
300000 "cg13581475"     "9.66e-01"      "6.27e-01"      "4.88e-01"      "-5.7782727"    "2.40e-02"      "63924286"      "MTHFD1'
InMemoryFileReader::Total number of lines data//input/gse59999/geo2r/Geo2r-gse5999-full-sign.txt =462177
Case: 99 .... 644
cg06410630 [571-5,0,0,20,90,2,12,] -- {RNF213;LOC100294362}, cg13560030 [442-59,12,19,37,4,27,] -- {NTN4}, cg02681173 [395-1,34
```

The file data/input/gse59999/geo2r/geo2RfeatureScoring.txt contains the result of this analysis.



```
Id  CPG Gene  Score Position
1 cg06410630  RNF213;LOC100294362 571 5,0,0,20,90,2,12,
2 cg13560030  NTN4  442 59,12,19,37,4,27,
3 cg02681173  LOC100190940  395 1,34,0,15,55,
4 cg09755579  SNORA70B;USP34  395 10,7,18,70,0,
5 cg20502977  COL6A3;COL6A3;COL6A3  357 1,39,3,0,
6 cg26124569  LPP;LPP;LPP 333 13,5,7,42,
7 cg24616138  CTBP2;CTBP2 313 4,12,1,70,
8 cg24584002  RNASEH1 294 19,17,39,31,
9 cg03946731  PKMYT1;PKMYT1 291 49,22,33,5,
10  cg20463995    281 38,43,0,38,
11  cg09618933    279 47,11,59,4,
12  cg10301401  LMF1  267 6,17,10,
13  cg08378782  RASGRP2;RASGRP2;RASGRP2 229 8,26,37,
14  cg21615831  KSR1  224 12,58,33,73,
15  cg07060505    221 0,10,69,
16  cg12176709  DRD2;DRD2 218 35,73,58,16,
17  cg02866639    212 27,34,27,
18  cg05897163  null  203 25,66,6,
```

This is summarized as shown below

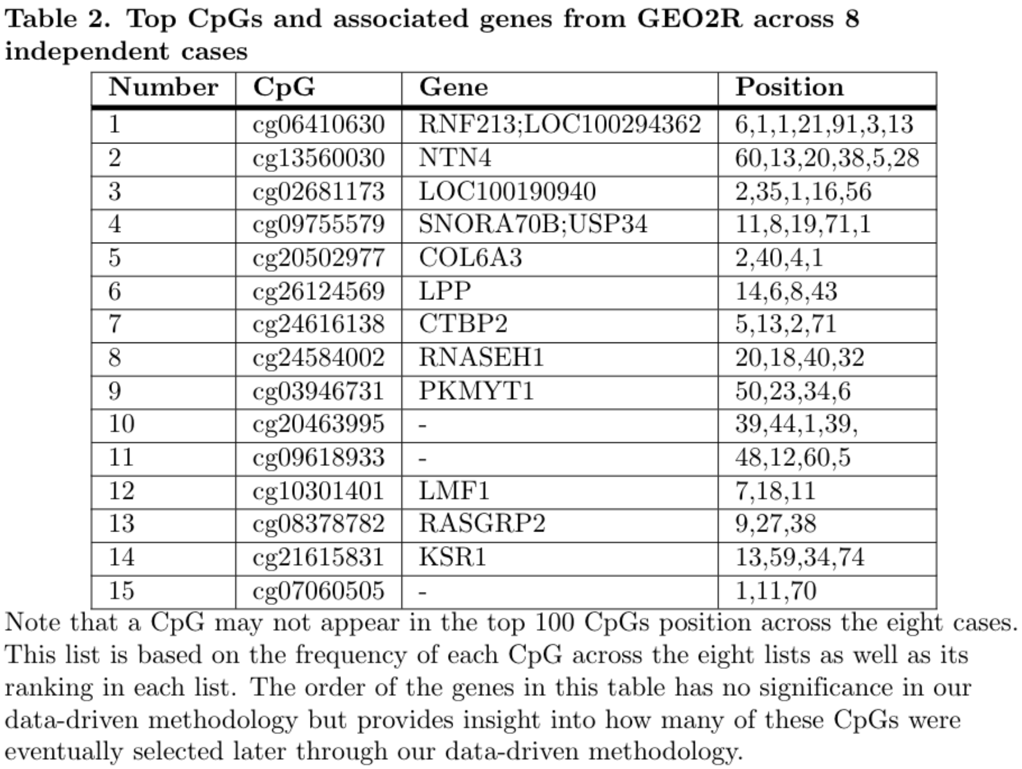


Table 2. Top CpGs and associated genes from GEO2R across 8 independent cases

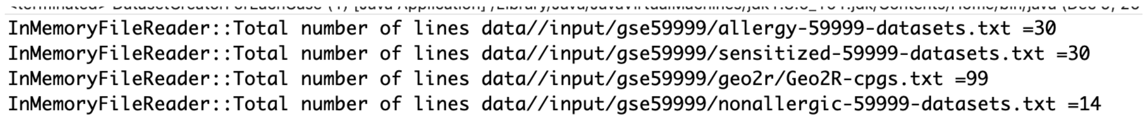
Number	CpG	Gene	Position
1	cg06410630	RNF213;LOC100294362	6,1,1,21,91,3,13
2	cg13560030	NTN4	60,13,20,38,5,28
3	cg02681173	LOC100190940	2,35,1,16,56
4	cg09755579	SNORA70B;USP34	11,8,19,71,1
5	cg20502977	COL6A3	2,40,4,1
6	cg26124569	LPP	14,6,8,43
7	cg24616138	CTBP2	5,13,2,71
8	cg24584002	RNASEH1	20,18,40,32
9	cg03946731	PKMYT1	50,23,34,6
10	cg20463995	-	39,44,1,39,
11	cg09618933	-	48,12,60,5
12	cg10301401	LMF1	7,18,11
13	cg08378782	RASGRP2	9,27,38
14	cg21615831	KSR1	13,59,34,74
15	cg07060505	-	1,11,70

Note that a CpG may not appear in the top 100 CpGs position across the eight cases. This list is based on the frequency of each CpG across the eight lists as well as its ranking in each list. The order of the genes in this table has no significance in our data-driven methodology but provides insight into how many of these CpGs were eventually selected later through our data-driven methodology.

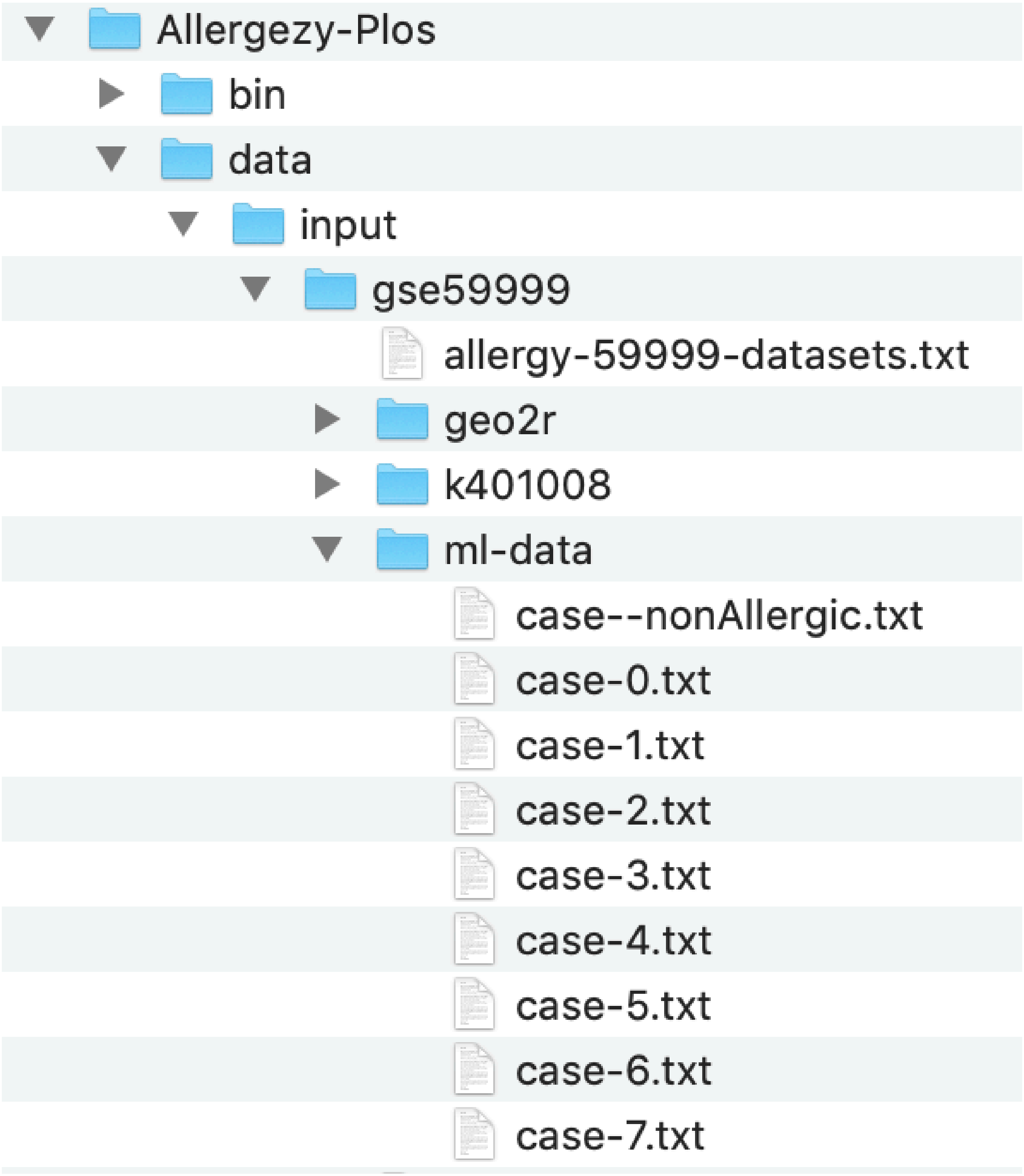
Note that this ranking really has no material significance in our methodology, since we evaluated all the unique 636 CpGs independently. It does provide some insights into features highlighted by GEO2R and those that appeared in our final CpG signature. The 20% overlap we found between the CpGs across the eight lists seems to suggest that the examples do have an effect on the CpGs that are differentially expressed between the two cohorts (when GEO2R is used). Thus, averaging the results across the eight independent runs, which is performed throughout the rest of the paper, should help in avoiding any potential biases, due to the distribution of the samples across the training, test, and hidden datasets.

6 Next, we will create smaller dataset files that contain the machine learning examples with 636 CpG values as columns.

com.allergezy.fa.dataset.DatasetCreatorForEachCase -- generates the file that has the examples for each case. The output from the program is shown below.



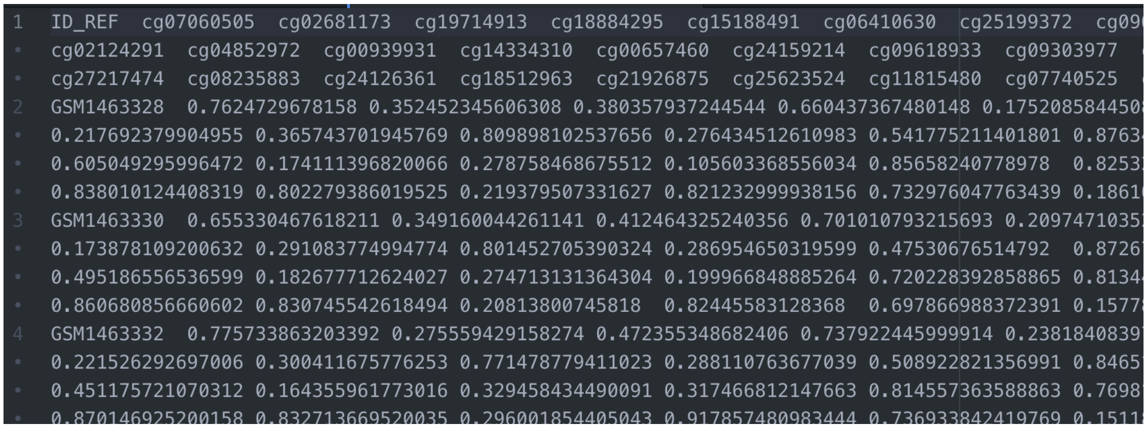
```
InMemoryFileReader::Total number of lines data//input/gse59999/allergy-59999-datasets.txt =30
InMemoryFileReader::Total number of lines data//input/gse59999/sensitized-59999-datasets.txt =30
InMemoryFileReader::Total number of lines data//input/gse59999/geo2r/Geo2R-cpgs.txt =99
InMemoryFileReader::Total number of lines data//input/gse59999/nonallergic-59999-datasets.txt =14
```



```
▼ Allergezy-Plos
    ▶ bin
    ▼ data
        ▼ input
            ▼ gse59999
                allergy-59999-datasets.txt
                ▶ geo2r
                ▶ k401008
                ▼ ml-data
                    case--nonAllergic.txt
                    case-0.txt
                    case-1.txt
                    case-2.txt
                    case-3.txt
                    case-4.txt
                    case-5.txt
                    case-6.txt
                    case-7.txt
```

In the input/gse59999/ml-data there are eight files case-0.txt, case-1.txt, and so on for the eight folds.

Attached is a screen shot of a part of the first file case-0.txt.



```
ID_REF  cg07060505  cg02681173  cg19714913  cg18884295  cg15188491  cg06410630  cg25199372  cg09
cg02124291  cg04852972  cg00939931  cg14334310  cg00657460  cg24159214  cg09618933  cg09303977
cg27217474  cg08235883  cg24126361  cg18512963  cg21926875  cg25623524  cg11815480  cg07740525
GSM1463328  0.7624729678158 0.35245234560630 8 0.380357937244544 0.660437367480148 0.175208584450
0.217692379904955 0.365743701945769 0.809898102537656 0.276434512610983 0.541775211401801 0.8763
0.605049295996472 0.174111396820066 0.278758468675512 0.105603368556034 0.85658240778978  0.8253
0.838010124408319 0.802279386019525 0.219379507331627 0.821232999938156 0.732976047763439 0.1861
GSM1463330  0.655330467618211 0.349160044261141 0.412464325240356 0.701010793215693 0.2097471035
0.173878109200632 0.291083774994774 0.801452705390324 0.286954650319599 0.47530676514792  0.8726
0.495186556536599 0.182677712624027 0.274713131364304 0.199966848885264 0.720228392858865 0.8134
0.860680856660602 0.830745542618494 0.20813800745818  0.82445583128368  0.697866988372391 0.1577
GSM1463332  0.775733863203392 0.275559429158274 0.472355348682406 0.737922445999914 0.2381840839
0.221526292697006 0.300411675776253 0.771478779411023 0.288110763677039 0.508922821356991 0.8465
0.451175721070312 0.164355961773016 0.329458434490091 0.317466812147663 0.814557363588863 0.7698
0.870146925200158 0.832713669520035 0.296001854405043 0.917857480983444 0.736933842419769 0.1511
```

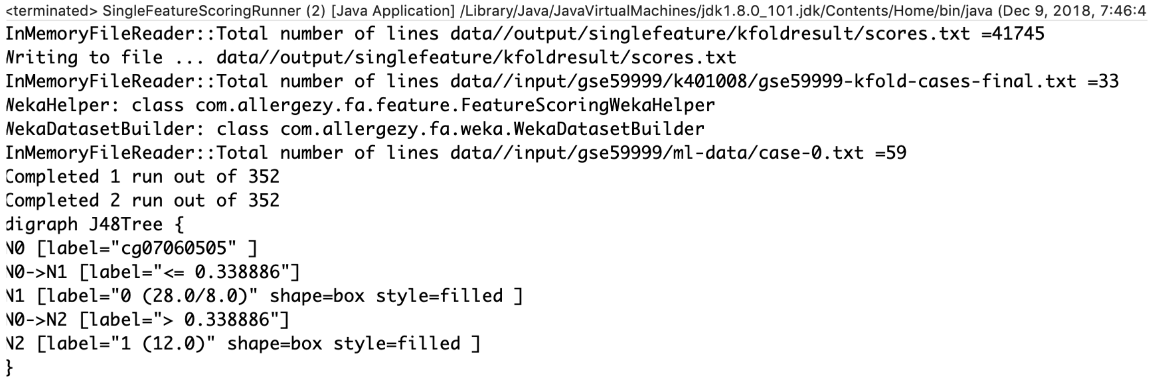
7 For each case, and for each of the 636 unique CpGs, we built four different machine learning models: Decision Trees (DT), Logistic Regression (LR), Radial Basis Function (RBF), and a Multi-Layer Perceptron (MLP), a deep learning network with two hidden layers of ten nodes each. Each predictive model only used one CpG and was built on the training data. For each case and each feature, the classifier with the highest testing accuracy was selected. Finally, the average hidden data (cross-validation set) accuracy across the eight independent cases was computed. 636 of these accuracy scores (one for each CpG) and 20,352 models (8 independent cases x 636 features x 4 classifiers) were created in total using this process, for the single input case.

com.allergezy.fa.feature.SingleFeatureScoringRunner is the class that runs this computation. A few class to note

- com.allergezy.fa.weka.WekaHelper interfaces with the Weka APIs for creating models
- Logic for writing details as part of the learning are SummaryRunResultWriterImpl -- summary of results

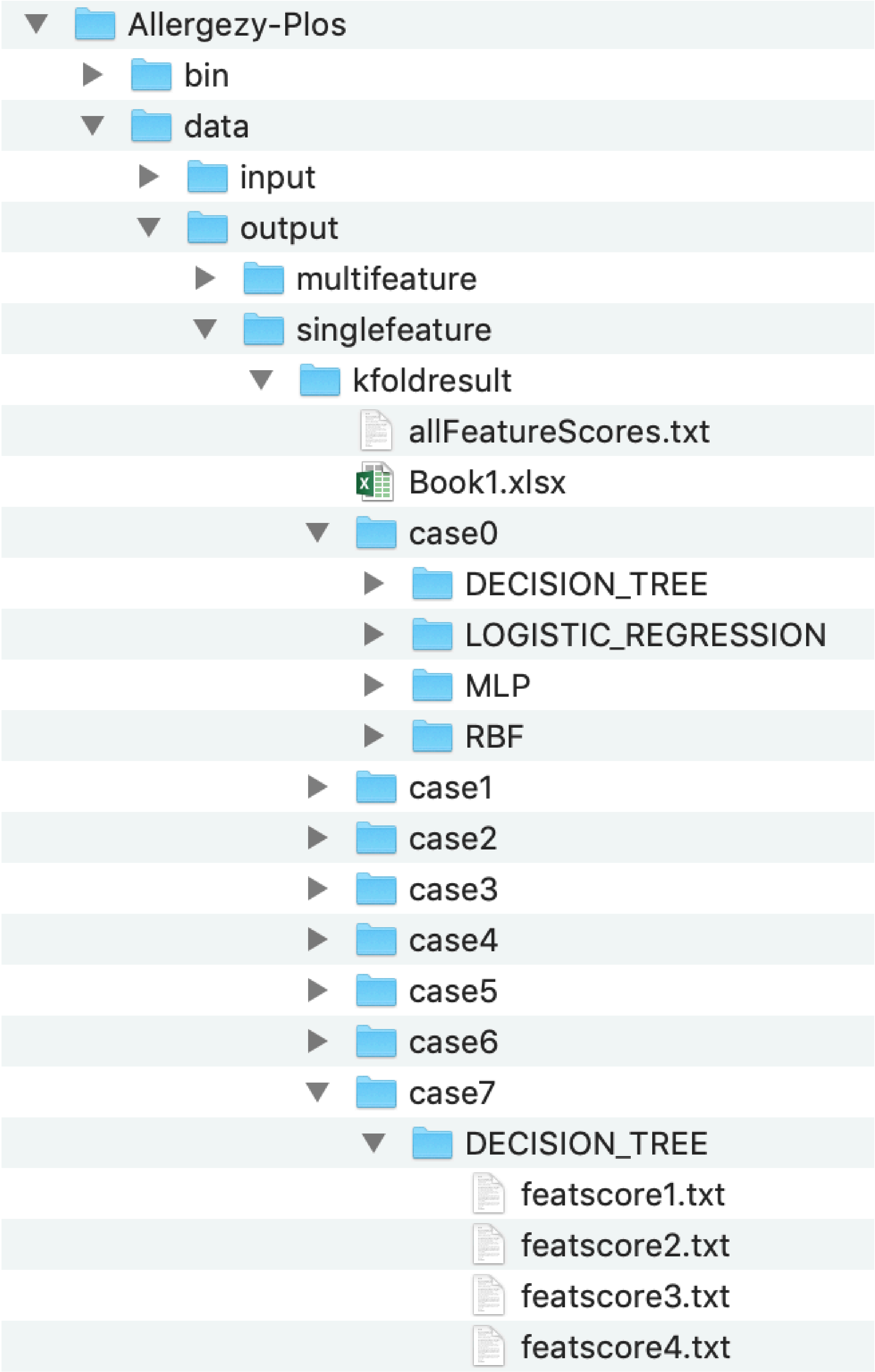
DetailedRunResultWriterImpl -- detailed output of each model,
BestModelForCasesResultWriterImpl - best model.

Run the main method com.allergezy.fa.feature.SingleFeatureScoringRunner. You should see an output similar to



```
<terminated> SingleFeatureScoringRunner (2) [Java Application] /Library/Java/JavaVirtualMachines/jdk1.8.0_101.jdk/Contents/Home/bin/java (Dec 9, 2018, 7:46:4
InMemoryFileReader::Total number of lines data//output/singlefeature/kfoldresult/scores.txt =41745
Writing to file ... data//output/singlefeature/kfoldresult/scores.txt
InMemoryFileReader::Total number of lines data//input/gse59999/k401008/gse59999-kfold-cases-final.txt =33
WekaHelper: class com.allergezy.fa.feature.FeatureScoringWekaHelper
WekaDatasetBuilder: class com.allergezy.fa.weka.WekaDatasetBuilder
InMemoryFileReader::Total number of lines data//input/gse59999/ml-data/case-0.txt =59
Completed 1 run out of 352
Completed 2 run out of 352
digraph J48Tree {
N0 [label="cg07060505" ]
N0->N1 [label="<= 0.338886"]
N1 [label="0 (28.0/8.0)" shape=box style=filled ]
N0->N2 [label="> 0.338886"]
N2 [label="1 (12.0)" shape=box style=filled ]
}
```

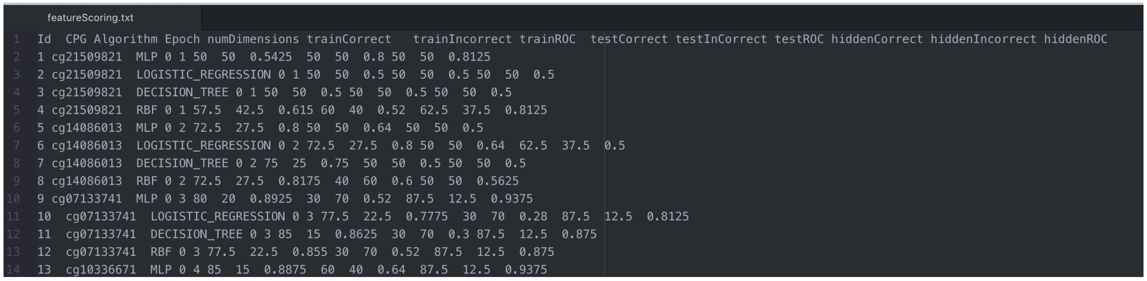
This program writes output to data/output/singlefeature/kfoldresult. Subdirectories case0, case1, ..., case7 has results for the eight folds. Within each directory are four sub directories: DECISION_TREE, LOGISTIC_REGRESSION, MLP, RBF. Within each directory are detailed results for each model and results -- featscore0.txt, featscore1.txt, ...,



```
▼ Allergezy-Plos
    ▶ bin
    ▼ data
        ▶ input
        ▼ output
            ▶ multifeature
            ▼ singlefeature
                ▼ kfoldresult
                      allFeatureScores.txt
                      Book1.xlsx
                    ▼ case0
                        ▶ DECISION_TREE
                        ▶ LOGISTIC_REGRESSION
                        ▶ MLP
                        ▶ RBF
                    ▶ case1
                    ▶ case2
                    ▶ case3
                    ▶ case4
                    ▶ case5
                    ▶ case6
                    ▼ case7
                        ▼ DECISION_TREE
                              featscore1.txt
                              featscore2.txt
                              featscore3.txt
                              featscore4.txt
```

The class com.allergezy.fa.feature.BestFeatureIdentifierUsingHiddenTestScores computes the average score for each feature across the eight cases.

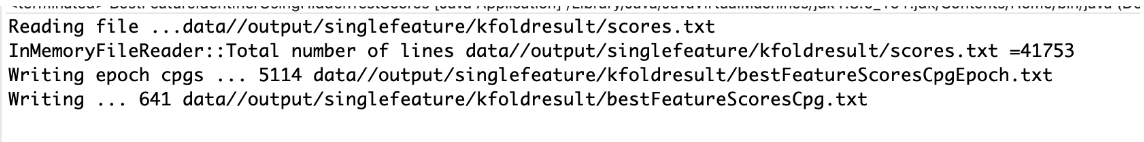
The following is a screen shot of the output file featureScoring.txt



```
featureScoring.txt
Id  CPG Algorithm Epoch numDimensions trainCorrect   trainIncorrect trainROC  testCorrect testInCorrect testROC hiddenCorrect hiddenIncorrect hiddenROC
1 cg21509821  MLP 0 1 50  50  0.5425  50  50  0.8 50  50  0.8125
2 cg21509821  LOGISTIC_REGRESSION 0 1 50  50  0.5 50  50  0.5 50  50  0.5
3 cg21509821  DECISION_TREE 0 1 50  50  0.5 50  50  0.5 50  50  0.5
4 cg21509821  RBF 0 1 57.5  42.5  0.615 60  40  0.52  62.5  37.5  0.8125
5 cg14086013  MLP 0 2 72.5  27.5  0.8 50  50  0.64  50  50  0.5
6 cg14086013  LOGISTIC_REGRESSION 0 2 72.5  27.5  0.8 50  50  0.64  62.5  37.5  0.5
7 cg14086013  DECISION_TREE 0 2 75  25  0.75  50  50  0.5 50  50  0.5
8 cg14086013  RBF 0 2 72.5  27.5  0.8175  40  60  0.6 50  50  0.5625
9 cg07133741  MLP 0 3 80  20  0.8925  30  70  0.52  87.5  12.5  0.9375
10  cg07133741  LOGISTIC_REGRESSION 0 3 77.5  22.5  0.7775  30  70  0.28  87.5  12.5  0.8125
11  cg07133741  DECISION_TREE 0 3 85  15  0.8625  30  70  0.3 87.5  12.5  0.875
12  cg07133741  RBF 0 3 77.5  22.5  0.855 30  70  0.52  87.5  12.5  0.875
13  cg10336671  MLP 0 4 85  15  0.8875  60  40  0.64  87.5  12.5  0.9375
```

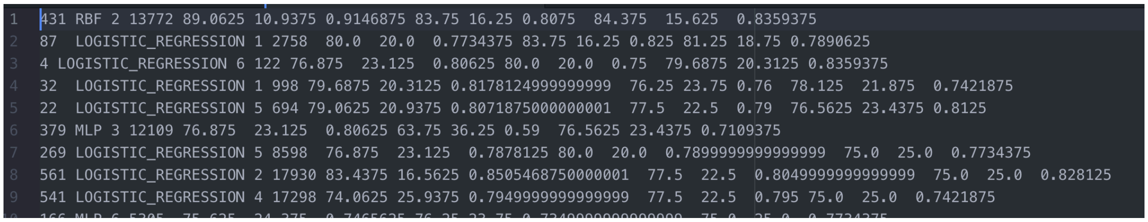
com.allergezy.fa.feature.BestFeatureIdentifierUsingHiddenTestScores

Output from the console



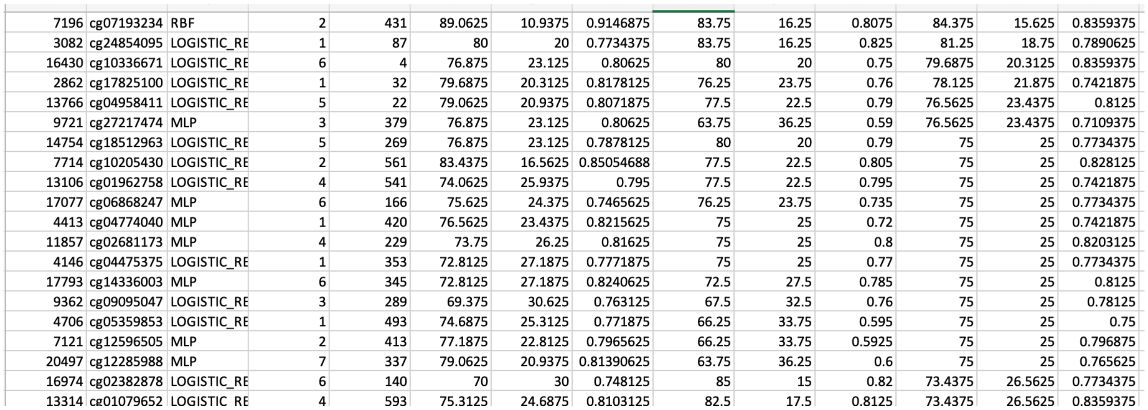
```
Reading file ...data//output/singlefeature/kfoldresult/scores.txt
InMemoryFileReader::Total number of lines data//output/singlefeature/kfoldresult/scores.txt =41753
Writing epoch cpgs ... 5114 data//output/singlefeature/kfoldresult/bestFeatureScoresCpgEpoch.txt
Writing ... 641 data//output/singlefeature/kfoldresult/bestFeatureScoresCpg.txt
```

Output file data/output/singlefeature/kfoldresult.bestFeatureScoresCpg.txt



```
431 RBF 2 13772 89.0625 10.9375 0.9146875 83.75 16.25 0.8075  84.375  15.625  0.8359375
87  LOGISTIC_REGRESSION 1 2758  80.0  20.0  0.7734375 83.75 16.25 0.825 81.25 18.75 0.7890625
4 LOGISTIC_REGRESSION 6 122 76.875  23.125  0.80625 80.0  20.0  0.75  79.6875 20.3125 0.8359375
32  LOGISTIC_REGRESSION 1 998 79.6875 20.3125 0.8178124999999999  76.25 23.75 0.76  78.125  21.875  0.7421875
22  LOGISTIC_REGRESSION 5 694 79.0625 20.9375 0.8071875000000001  77.5  22.5  0.79  76.5625 23.4375 0.8125
379 MLP 3 12109 76.875  23.125  0.80625 63.75 36.25 0.59  76.5625 23.4375 0.7109375
269 LOGISTIC_REGRESSION 5 8598  76.875  23.125  0.7878125 80.0  20.0  0.7899999999999999  75.0  25.0  0.7734375
561 LOGISTIC_REGRESSION 2 17930 83.4375 16.5625 0.8505468750000001  77.5  22.5  0.8049999999999999  75.0  25.0  0.828125
541 LOGISTIC_REGRESSION 4 17298 74.0625 25.9375 0.7949999999999999  77.5  22.5  0.795 75.0  25.0  0.7421875
166 MLP 6 5305  75.625  24.375  0.7465625 76.25 23.75 0.7349999999999999  75.0  25.0  0.7734375
```

Attached is the screenshot of the final file, sorted by average hidden test accuracy.



7196	cg07193234	RBF	2	431	89.0625	10.9375	0.9146875	83.75	16.25	0.8075	84.375	15.625	0.8359375
3082	cg24854095	LOGISTIC_RE	1	87	80	20	0.7734375	83.75	16.25	0.825	81.25	18.75	0.7890625
16430	cg10336671	LOGISTIC_RE	6	4	76.875	23.125	0.80625	80	20	0.75	79.6875	20.3125	0.8359375
2862	cg17825100	LOGISTIC_RE	1	32	79.6875	20.3125	0.8178125	76.25	23.75	0.76	78.125	21.875	0.7421875
13766	cg04958411	LOGISTIC_RE	5	22	79.0625	20.9375	0.8071875	77.5	22.5	0.79	76.5625	23.4375	0.8125
9721	cg27217474	MLP	3	379	76.875	23.125	0.80625	63.75	36.25	0.59	76.5625	23.4375	0.7109375
14754	cg18512963	LOGISTIC_RE	5	269	76.875	23.125	0.7878125	80	20	0.79	75	25	0.7734375
7714	cg10205430	LOGISTIC_RE	2	561	83.4375	16.5625	0.85054688	77.5	22.5	0.805	75	25	0.828125
13106	cg01962758	LOGISTIC_RE	4	541	74.0625	25.9375	0.795	77.5	22.5	0.795	75	25	0.7421875
17077	cg06868247	MLP	6	166	75.625	24.375	0.7465625	76.25	23.75	0.735	75	25	0.7734375
4413	cg04774040	MLP	1	420	76.5625	23.4375	0.8215625	75	25	0.72	75	25	0.7421875
11857	cg02681173	MLP	4	229	73.75	26.25	0.81625	75	25	0.8	75	25	0.8203125
4146	cg04475375	LOGISTIC_RE	1	353	72.8125	27.1875	0.7771875	75	25	0.77	75	25	0.7734375
17793	cg14336003	MLP	6	345	72.8125	27.1875	0.8240625	72.5	27.5	0.785	75	25	0.8125
9362	cg09095047	LOGISTIC_RE	3	289	69.375	30.625	0.763125	67.5	32.5	0.76	75	25	0.78125
4706	cg05359853	LOGISTIC_RE	1	493	74.6875	25.3125	0.771875	66.25	33.75	0.595	75	25	0.75
7121	cg12596505	MLP	2	413	77.1875	22.8125	0.7965625	66.25	33.75	0.5925	75	25	0.796875
20497	cg12285988	MLP	7	337	79.0625	20.9375	0.81390625	63.75	36.25	0.6	75	25	0.765625
16974	cg02382878	LOGISTIC_RE	6	140	70	30	0.748125	85	15	0.82	73.4375	26.5625	0.7734375
13314	cg01079652	LOGISTIC_RE	4	593	75.3125	24.6875	0.8103125	82.5	17.5	0.8125	73.4375	26.5625	0.8359375

The following graph summarizes the results.

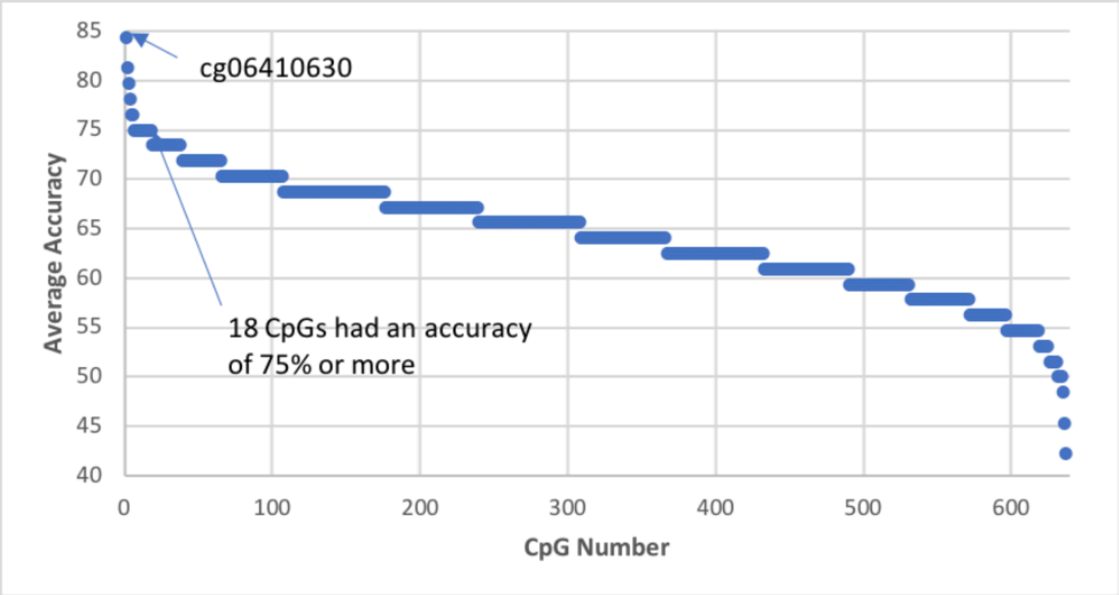


Figure 1. Average accuracy across eight independent cases for singular CpG features. Each CpG was used as an input to create classifiers using four different machine learning methods. The classifier with the best score on test data was then selected and evaluated on the hidden dataset. The above accuracy for each CpG is the average accuracy across the eight different cases. cg06410630 was the top CpG with an average accuracy of 84.375%. 18 CpGs had a score of 75% or more.

The top CpGs with associated genes are

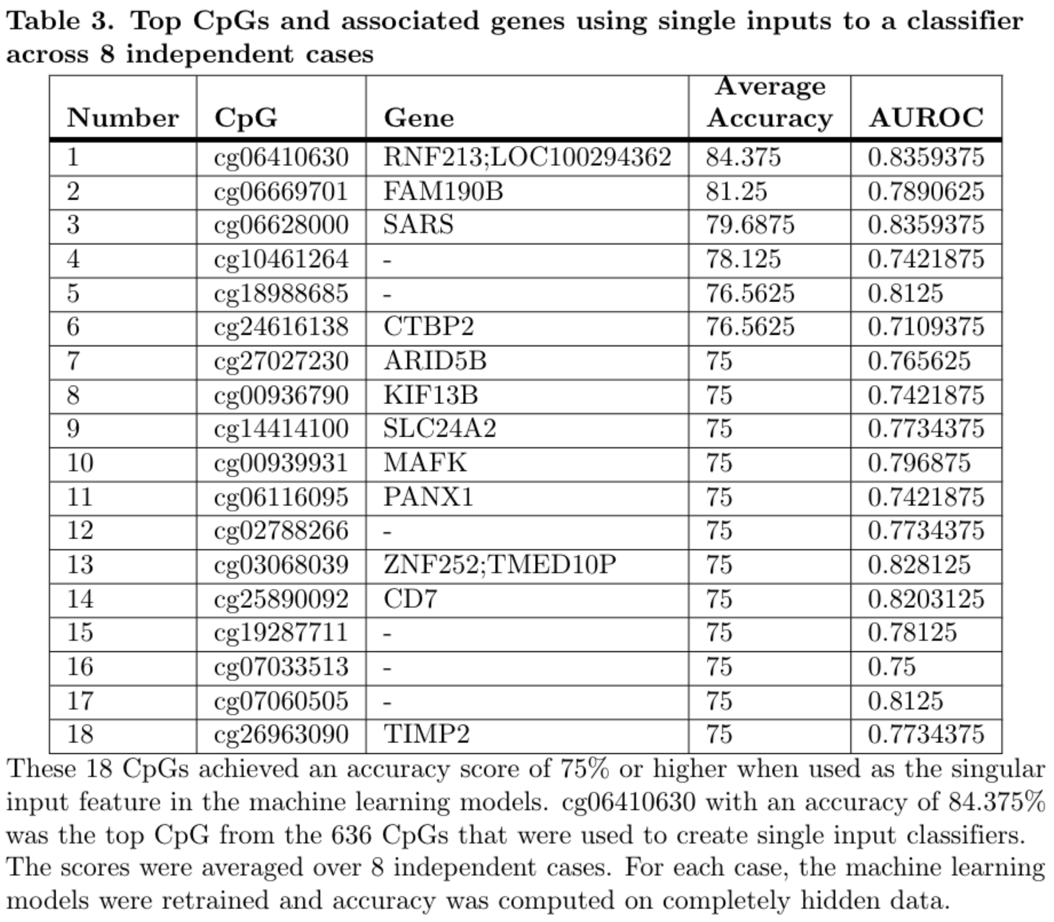


Table 3. Top CpGs and associated genes using single inputs to a classifier across 8 independent cases

Number	CpG	Gene	Average Accuracy	AUROC
1	cg06410630	RNF213;LOC100294362	84.375	0.8359375
2	cg06669701	FAM190B	81.25	0.7890625
3	cg06628000	SARS	79.6875	0.8359375
4	cg10461264	-	78.125	0.7421875
5	cg18988685	-	76.5625	0.8125
6	cg24616138	CTBP2	76.5625	0.7109375
7	cg27027230	ARID5B	75	0.765625
8	cg00936790	KIF13B	75	0.7421875
9	cg14414100	SLC24A2	75	0.7734375
10	cg00939931	MAFK	75	0.796875
11	cg06116095	PANX1	75	0.7421875
12	cg02788266	-	75	0.7734375
13	cg03068039	ZNF252;TMED10P	75	0.828125
14	cg25890092	CD7	75	0.8203125
15	cg19287711	-	75	0.78125
16	cg07033513	-	75	0.75
17	cg07060505	-	75	0.8125
18	cg26963090	TIMP2	75	0.7734375

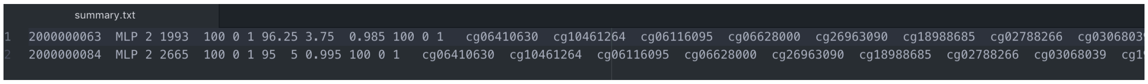
These 18 CpGs achieved an accuracy score of 75% or higher when used as the singular input feature in the machine learning models. cg06410630 with an accuracy of 84.375% was the top CpG from the 636 CpGs that were used to create single input classifiers. The scores were averaged over 8 independent cases. For each case, the machine learning models were retrained and accuracy was computed on completely hidden data.

8 We sought to increase the number of features used by the classifier while still avoiding over-fitting and encouraging generalization. Therefore, we selected the first eighteen CpGs, based on the accuracy scores computed above, and combined them two at a time, followed by three at a time, and so on until combinations of twelve were reached. Given the large number of potential combinations, each model was limited to a small subset of strong CpG-lists, to which a new input feature was added. On an average, we tried about 200 unique combinations for a given number of input features. Again, each unique input feature combination set was run 4 × 8 = 32 times, to account for the four different classifier methods and 8 independent sample-distribution "cases".

The class com.allergezy.fa.multifeature.casegenerator.MultiFeatureCasesGenerator is used to create next set of cases to be tried.

9 For a given number of input features, the models were ranked using their average accuracy scores on hidden data cohorts (described above). Odd numbers of models, starting from 1 to 101, were combined together using a simple voting scheme, i.e., each model independently predicted whether a sample was classified as FA or sensitized and the final prediction was the majority of predictions made across the different models.

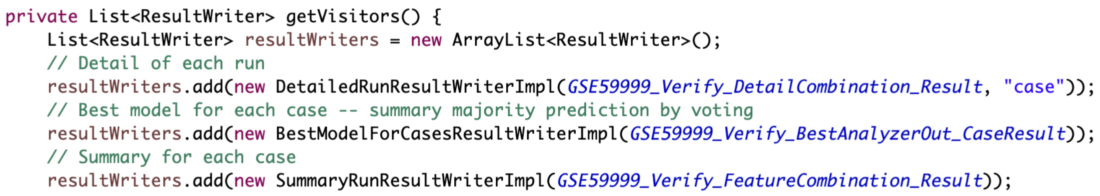
The class com.allergezy.fa.kfold.verify.VerifyWekaDatasetRunner runs the set of cases specified in the data/output/multifeature/setup/summary.txt file. A screen shot of an example file is shown next.



summary.txt

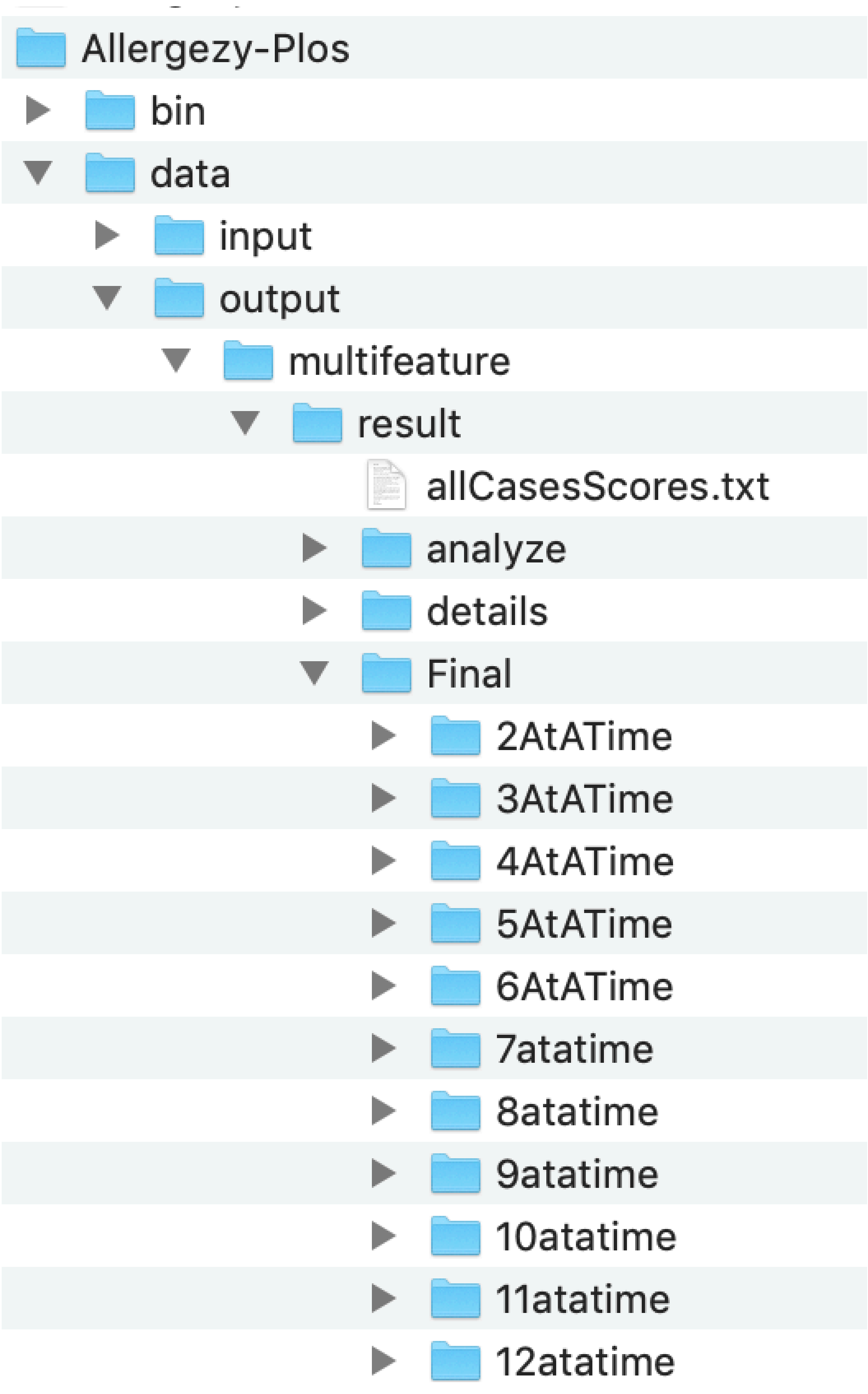
```
2000000063  MLP 2 1993  100 0 1 96.25 3.75  0.985 100 0 1   cg06410630  cg10461264  cg06116095  cg06628000  cg26963090  cg18988685  cg02788266  cg0306803
2000000084  MLP 2 2665  100 0 1 95  5 0.995 100 0 1   cg06410630  cg10461264  cg06116095  cg06628000  cg26963090  cg18988685  cg02788266  cg03068039  cg1
```

This class has a visitor BestModelForCasesResultWriterImpl configured to calculate majority prediction.



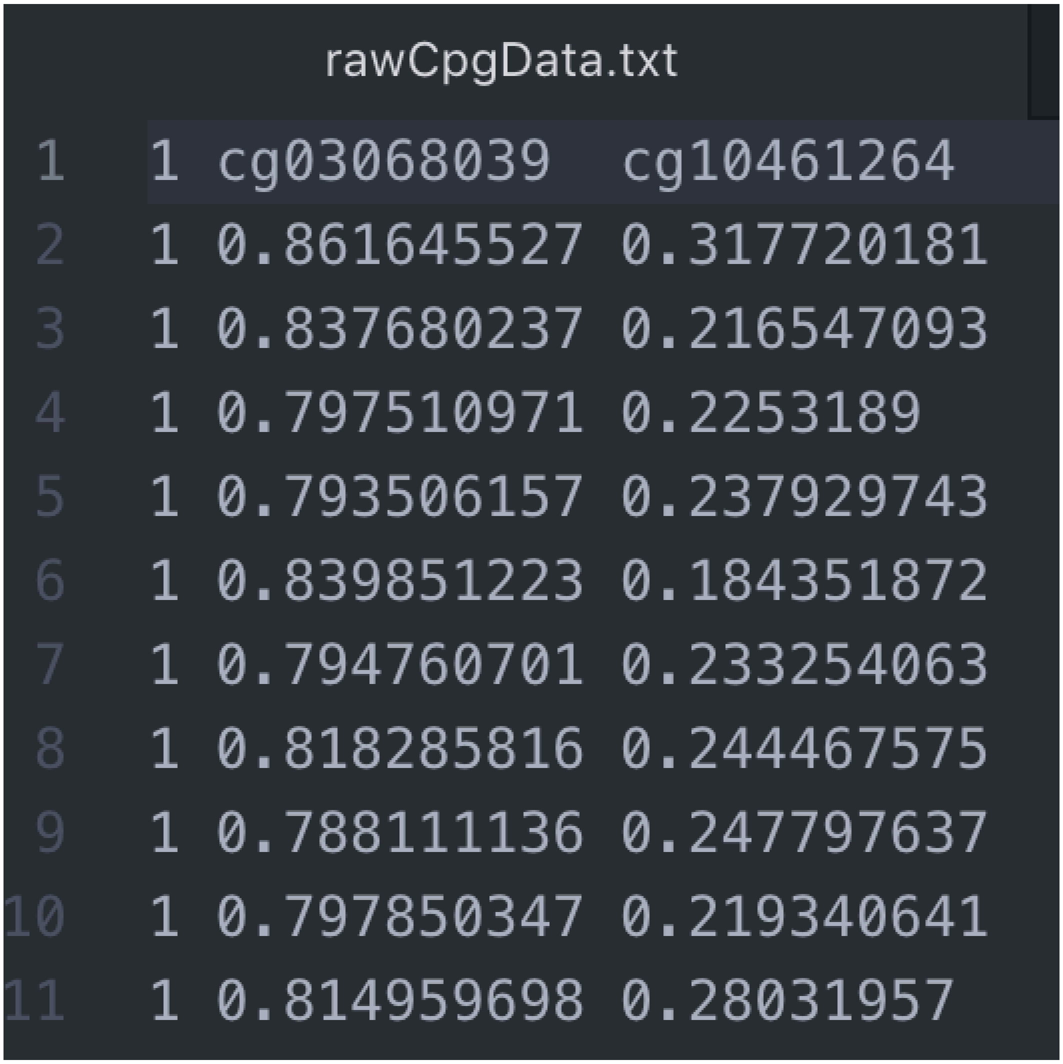
```java
private List<ResultWriter> getVisitors() {
    List<ResultWriter> resultWriters = new ArrayList<ResultWriter>();
    // Detail of each run
    resultWriters.add(new DetailedRunResultWriterImpl(GSE59999_Verify_DetailCombination_Result, "case"));
    // Best model for each case -- summary majority prediction by voting
    resultWriters.add(new BestModelForCasesResultWriterImpl(GSE59999_Verify_BestAnalyzerOut_CaseResult));
    // Summary for each case
    resultWriters.add(new SummaryRunResultWriterImpl(GSE59999_Verify_FeatureCombination_Result));
```

Iterate different combinations of CpG features using the summary.txt file as input, run the results, select the best models, and try new combinations adding one feature at a time. The results for various combinations are in the directory data/output/multfeature/result/Final.



```
Allergezy-Plos
  ▶ bin
  ▼ data
      ▶ input
      ▼ output
          ▼ multifeature
              ▼ result
                    allCasesScores.txt
                  ▶ analyze
                  ▶ details
                  ▼ Final
                      ▶ 2AtATime
                      ▶ 3AtATime
                      ▶ 4AtATime
                      ▶ 5AtATime
                      ▶ 6AtATime
                      ▶ 7atatime
                      ▶ 8atatime
                      ▶ 9atatime
                      ▶ 10atatime
                      ▶ 11atatime
                      ▶ 12atatime
```

10 The file com.allergezy.fa.rawdata.RawCpGDataExtractor can extract raw CpG values for a CpG across the samples and can be used for generating data plots. Here is an example file.



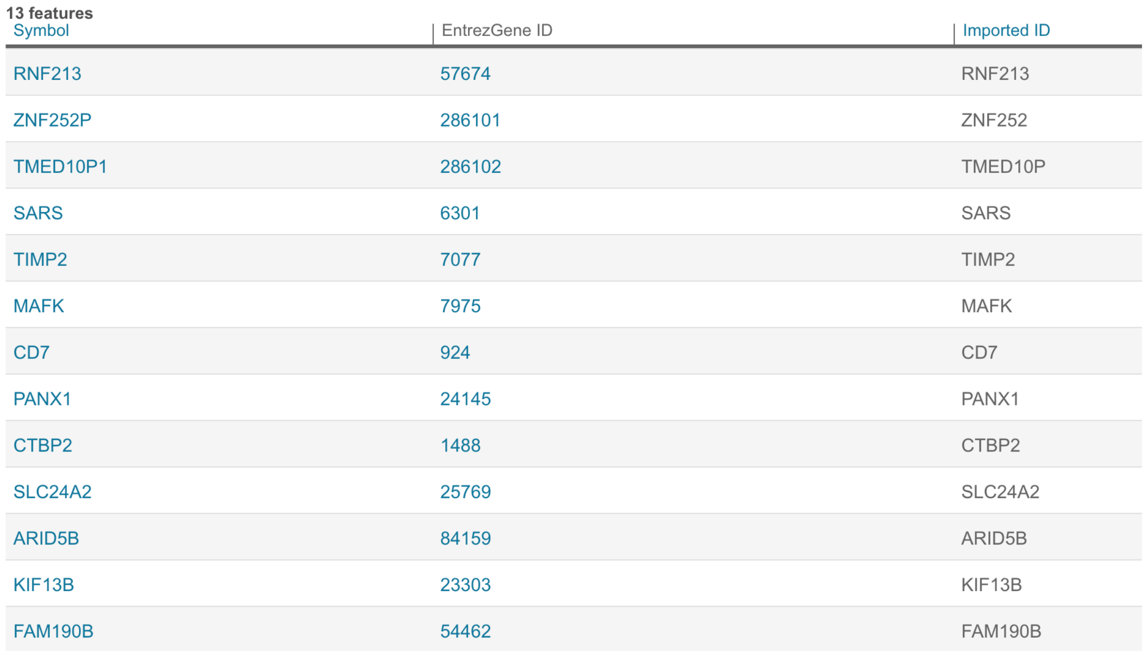
rawCpgData.txt

```
1 cg03068039  cg10461264
1 0.861645527 0.317720181
1 0.837680237 0.216547093
1 0.797510971 0.2253189
1 0.793506157 0.237929743
1 0.839851223 0.184351872
1 0.794760701 0.233254063
1 0.818285816 0.244467575
1 0.788111136 0.247797637
1 0.797850347 0.219340641
1 0.814959698 0.28031957
```

com.allergezy.fa.featureanalyzer.CpGFrequencyFeatureAnalyzer
Analyzes the frequency of various CpG across the summary file

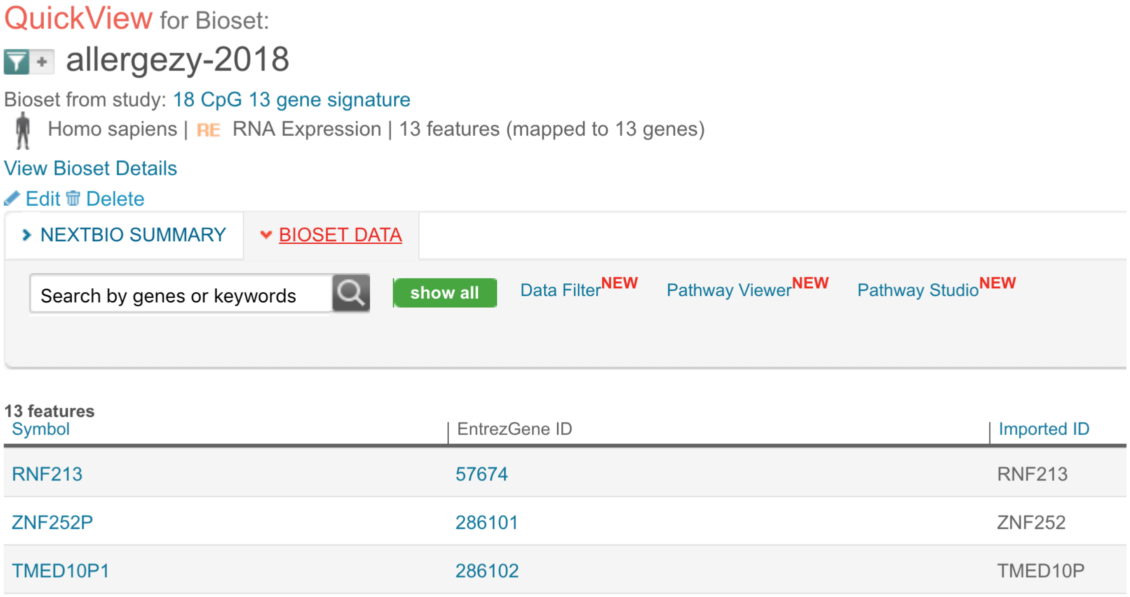
11 Gene set analysis can provide biological context, as well as insights into disease mechanisms and possible treatments. Biological enrichment was performed by applying Illumina’s BaseSpace Correlation Engine to the 13-gene list. To understand these genes better we identified tissues where the genes were expressed, found associated biological pathways, used gene ontology concepts to identify functionally-related gene sets, connected the 13-gene list to the Broad positional gene sets, and connected the gene signature to protein families.

The 13-gene signature was imported into Illumina BaseSpace Correlation Engine



13 features

Symbol	EntrezGene ID	Imported ID
RNF213	57674	RNF213
ZNF252P	286101	ZNF252
TMED10P1	286102	TMED10P
SARS	6301	SARS
TIMP2	7077	TIMP2
MAFK	7975	MAFK
CD7	924	CD7
PANX1	24145	PANX1
CTBP2	1488	CTBP2
SLC24A2	25769	SLC24A2
ARID5B	84159	ARID5B
KIF13B	23303	KIF13B
FAM190B	54462	FAM190B



QuickView for Bioset:

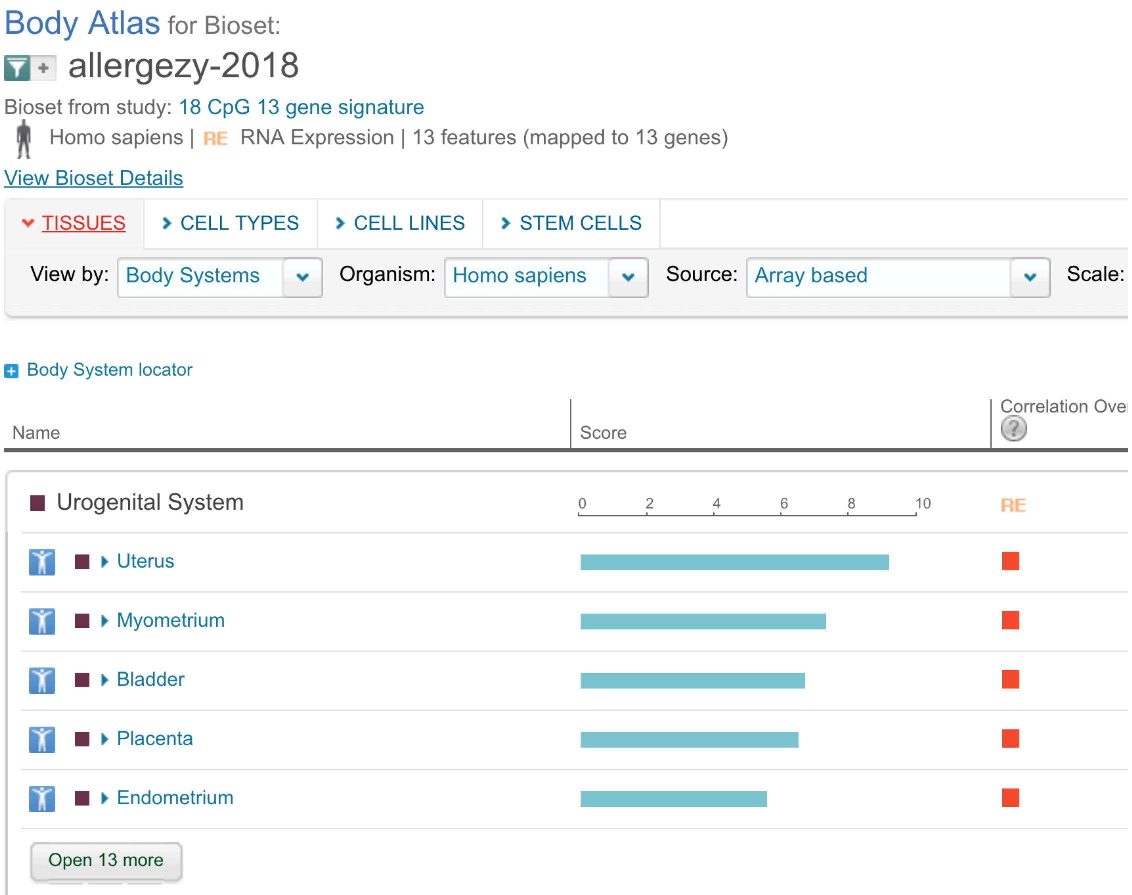
allergezy-2018

Bioset from study: 18 CpG 13 gene signature

Homo sapiens | RE RNA Expression | 13 features (mapped to 13 genes)

13 features

Symbol	EntrezGene ID	Imported ID
RNF213	57674	RNF213
ZNF252P	286101	ZNF252
TMED10P1	286102	TMED10P



Body Atlas for Bioset:

allergezy-2018

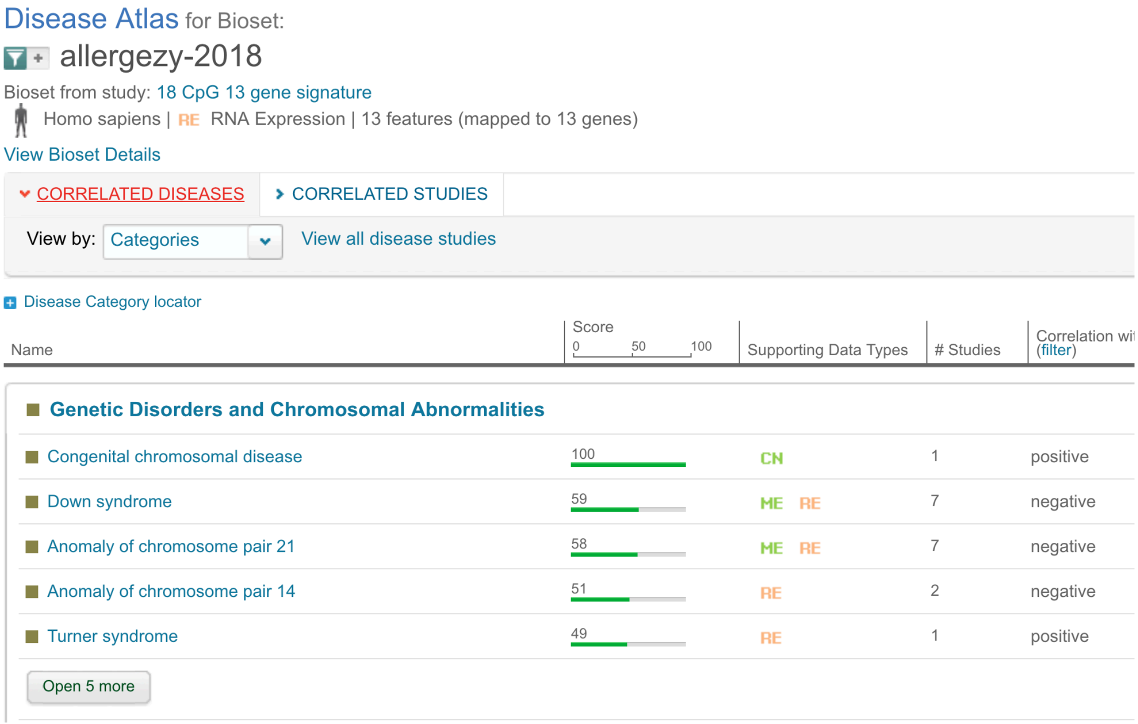
Bioset from study: 18 CpG 13 gene signature

Homo sapiens | RE RNA Expression | 13 features (mapped to 13 genes)

View Bioset Details

View by: Body Systems Organism: Homo sapiens Source: Array based Scale:

Name	Score	Correlation Ove
Urogenital System		RE
Uterus		🟥
Myometrium		🟥
Bladder		🟥
Placenta		🟥
Endometrium		🟥



Disease Atlas for Bioset:

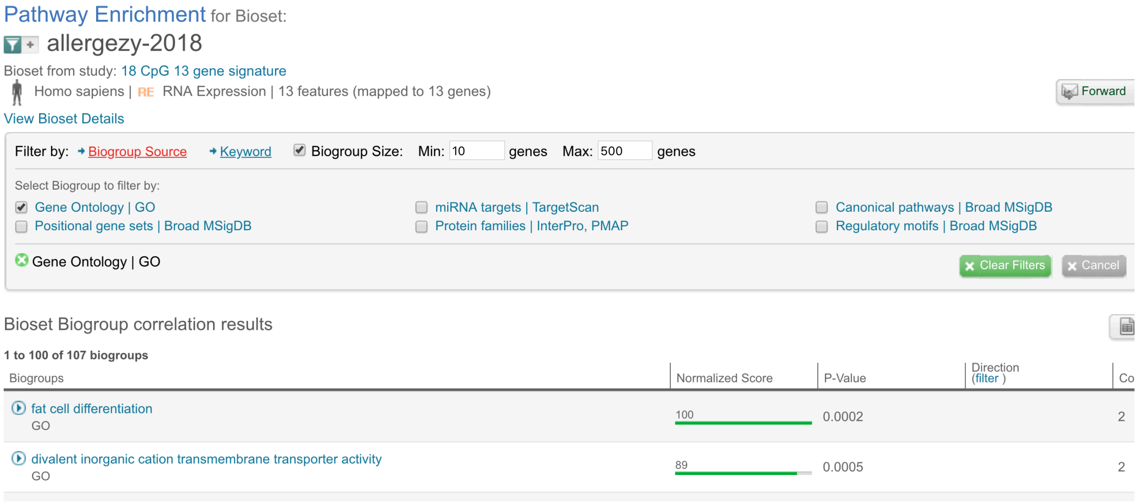
allergezy-2018

Bioset from study: 18 CpG 13 gene signature

Homo sapiens | RE RNA Expression | 13 features (mapped to 13 genes)

View Bioset Details

Name	Score 0 50 100	Supporting Data Types	# Studies	Correlation wi (filter)
Genetic Disorders and Chromosomal Abnormalities				
Congenital chromosomal disease	100	CN	1	positive
Down syndrome	59	ME RE	7	negative
Anomaly of chromosome pair 21	58	ME RE	7	negative
Anomaly of chromosome pair 14	51	RE	2	negative
Turner syndrome	49	RE	1	positive



Pathway Enrichment for Bioset:

allergezy-2018

Bioset from study: 18 CpG 13 gene signature

Homo sapiens | RE RNA Expression | 13 features (mapped to 13 genes)

View Bioset Details

Filter by: Biogroup Source Keyword ☑ Biogroup Size: Min: 10 genes Max: 500 genes

Select Biogroup to filter by:

☑ Gene Ontology | GO
☐ Positional gene sets | Broad MSigDB
☐ miRNA targets | TargetScan
☐ Protein families | InterPro, PMAP
☐ Canonical pathways | Broad MSigDB
☐ Regulatory motifs | Broad MSigDB

Gene Ontology | GO

Bioset Biogroup correlation results

1 to 100 of 107 biogroups

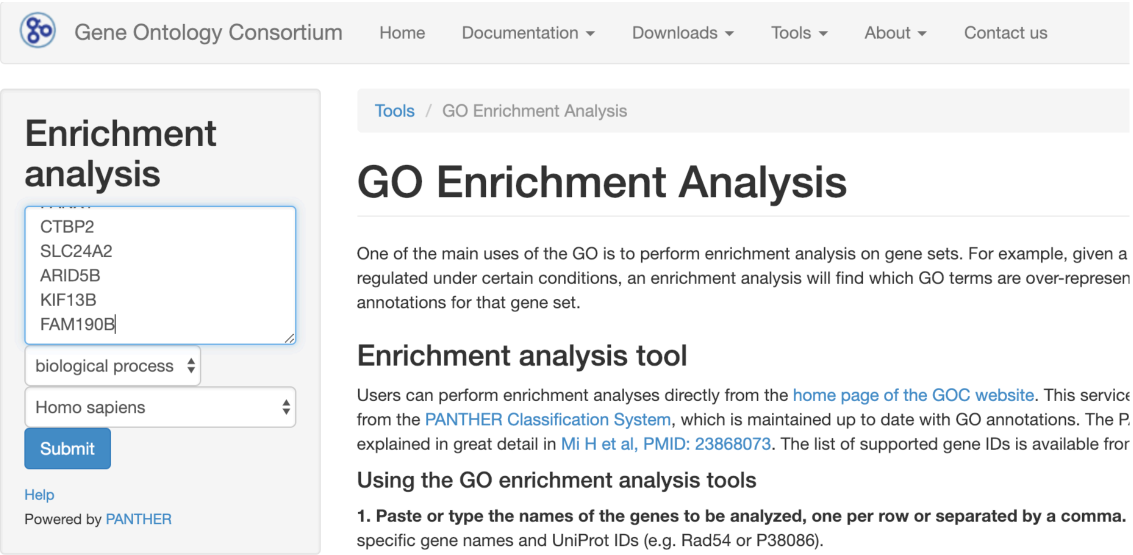
Biogroups	Normalized Score	P-Value	Direction (filter)	Co
fat cell differentiation GO	100	0.0002		2
divalent inorganic cation transmembrane transporter activity GO	89	0.0005		2

12 GO provides a framework for analyzing a group of co-expressed genes and provides insights into whether particular genes are involved in diseases. Similar to other ontologies, GO consists of terms and relationships represented in a directed acylic

graph. GO focuses on using the terms in the ontology to describe gene functions. In GO, gene function is classified along three categories: molecular functions, cellular components, and biological processes.

The GO Enrichment Analysis Tool was used to find GO terms associated with the 13-gene signature. Enter the following gene list as shown

RNF213
SARS
ZNF252
TMED10P
ABCF2
TIMP2
MAFK
CD7
PANX1
CTBP2
SLC24A2
ARID5B
KIF13B
FAM190B



Enrichment analysis

CTBP2
SLC24A2
ARID5B
KIF13B
FAM190B

biological process

Homo sapiens

GO Enrichment Analysis

One of the main uses of the GO is to perform enrichment analysis on gene sets. For example, given a regulated under certain conditions, an enrichment analysis will find which GO terms are over-represen annotations for that gene set.

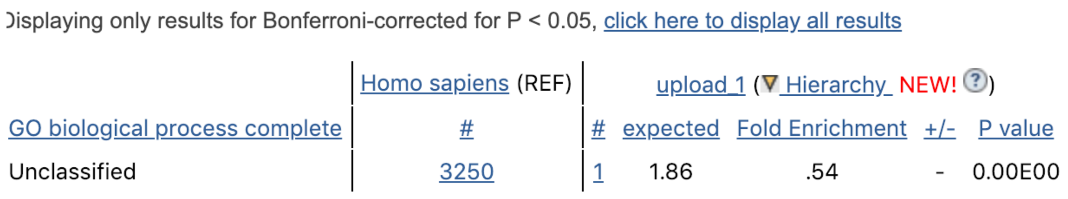
Enrichment analysis tool

Users can perform enrichment analyses directly from the home page of the GOC website. This service from the PANTHER Classification System, which is maintained up to date with GO annotations. The P explained in great detail in Mi H et al, PMID: 23868073. The list of supported gene IDs is available fror

Using the GO enrichment analysis tools

1. Paste or type the names of the genes to be analyzed, one per row or separated by a comma. specific gene names and UniProt IDs (e.g. Rad54 or P38086).

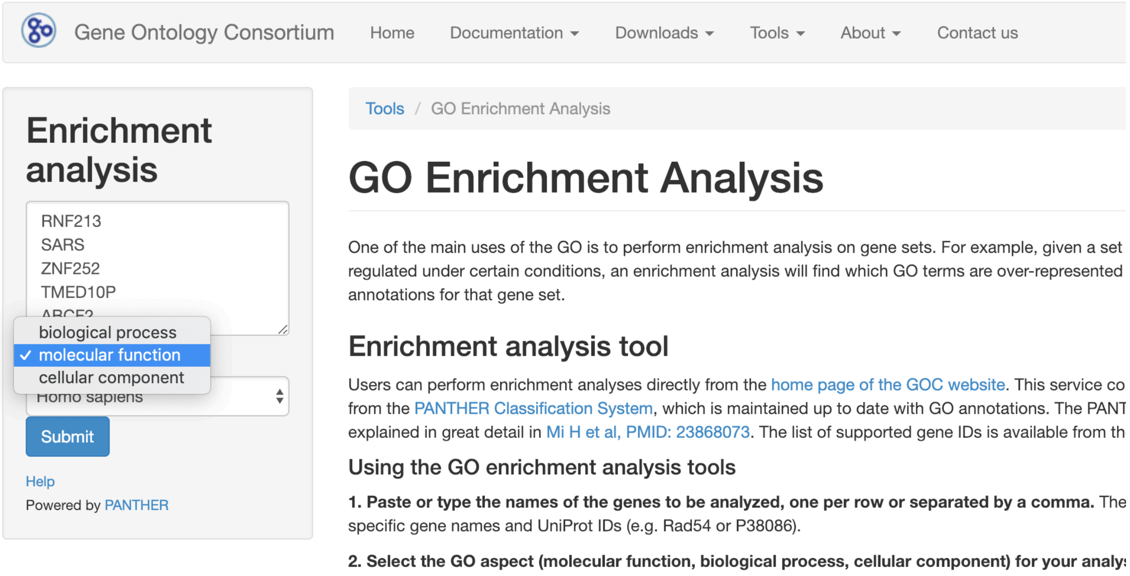
After computation the results are shown below.



Displaying only results for Bonferroni-corrected for P < 0.05, click here to display all results

	Homo sapiens (REF)	upload 1 (Hierarchy NEW! ?)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
Unclassified	3250	1	1.86	.54	-	0.00E00

Follow the same procedure for the other two options: molecular function and cellular component, as shown below.



Enrichment analysis

RNF213
SARS
ZNF252
TMED10P
ABCF2

Powered by PANTHER

GO Enrichment Analysis

One of the main uses of the GO is to perform enrichment analysis on gene sets. For example, given a set regulated under certain conditions, an enrichment analysis will find which GO terms are over-represented annotations for that gene set.

Enrichment analysis tool

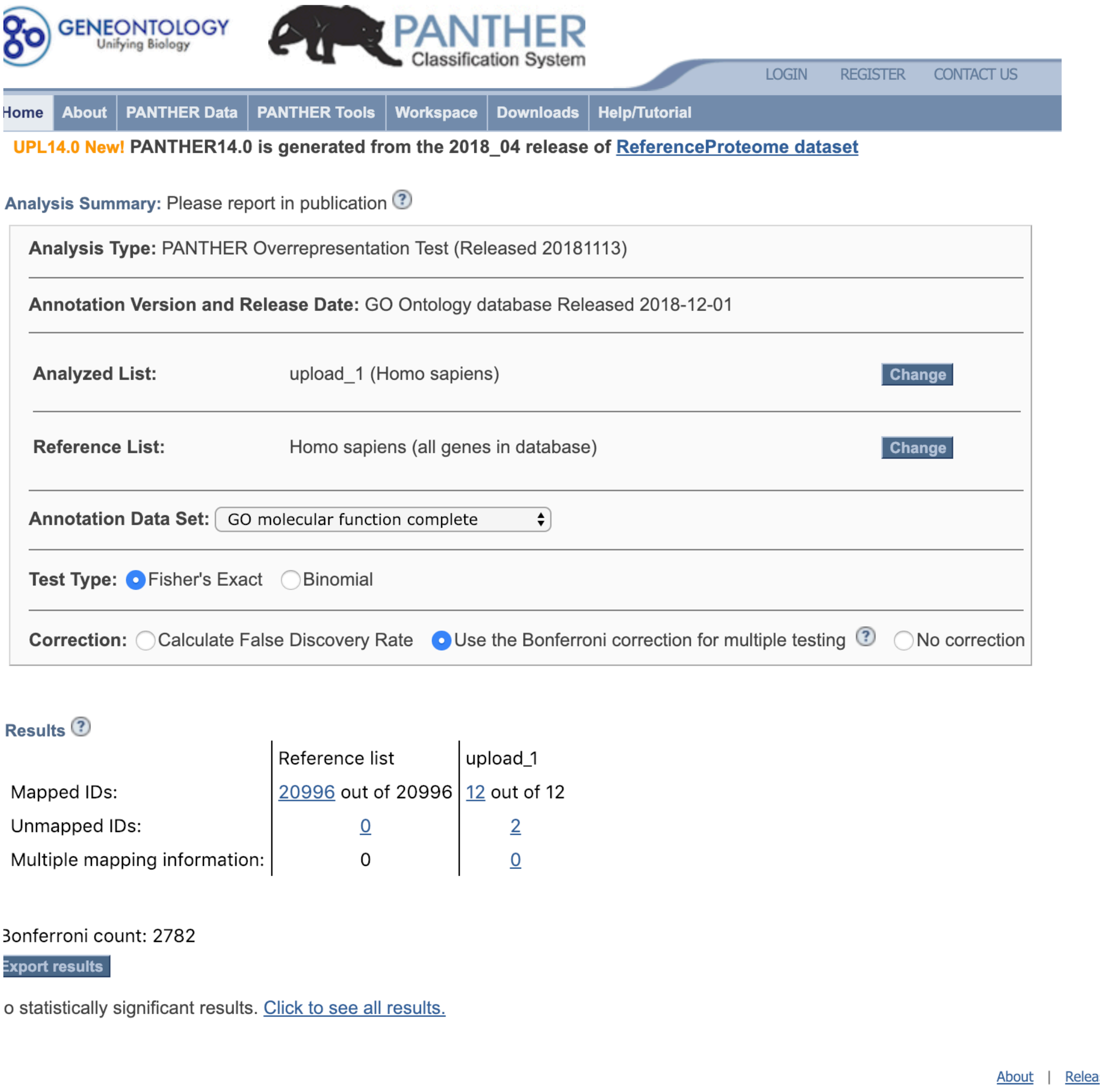
Users can perform enrichment analyses directly from the home page of the GOC website. This service co from the PANTHER Classification System, which is maintained up to date with GO annotations. The PANT explained in great detail in Mi H et al, PMID: 23868073. The list of supported gene IDs is available from th

Using the GO enrichment analysis tools

1. Paste or type the names of the genes to be analyzed, one per row or separated by a comma. The specific gene names and UniProt IDs (e.g. Rad54 or P38086).

2. Select the GO aspect (molecular function, biological process, cellular component) for your analy

The results from this analaysis is shown below.



UPL14.0 New! PANTHER14.0 is generated from the 2018_04 release of ReferenceProteome dataset

Analysis Summary: Please report in publication ?

Analysis Type: PANTHER Overrepresentation Test (Released 20181113)

Annotation Version and Release Date: GO Ontology database Released 2018-12-01

Analyzed List: upload_1 (Homo sapiens)

Reference List: Homo sapiens (all genes in database)

Annotation Data Set: GO molecular function complete

Test Type: ● Fisher's Exact ○ Binomial

Correction: ○ Calculate False Discovery Rate ● Use the Bonferroni correction for multiple testing ? ○ No correction

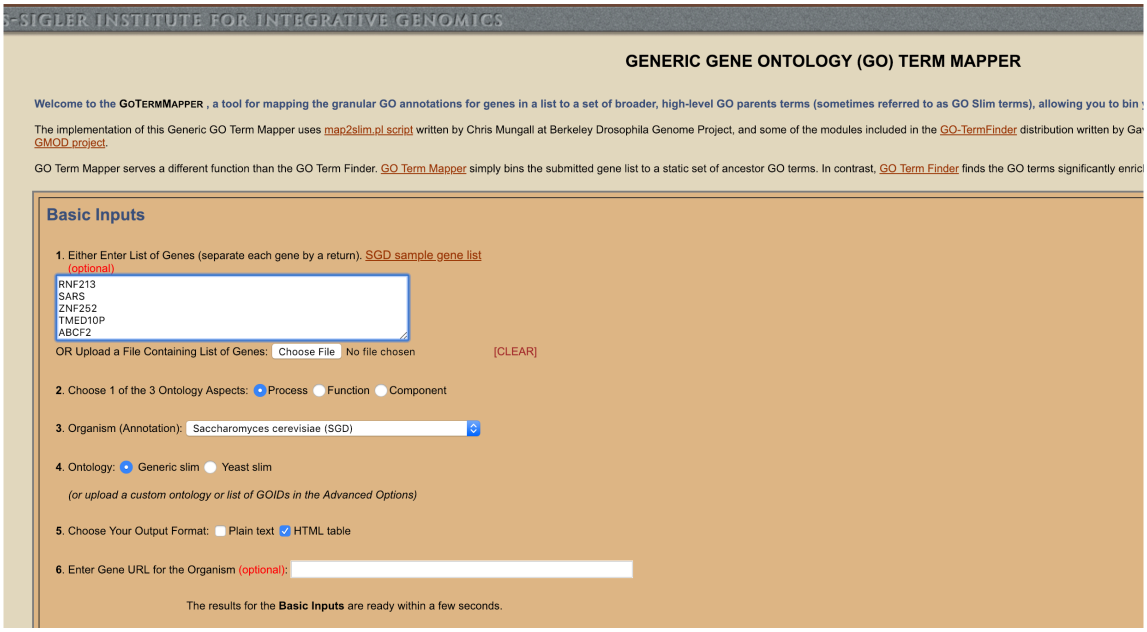
Results ?

	Reference list	upload_1
Mapped IDs:	20996 out of 20996	12 out of 12
Unmapped IDs:	0	2
Multiple mapping information:	0	0

Bonferroni count: 2782

o statistically significant results. Click to see all results.

Next, the **Generic Gene Ontology (GO) Term Mapper** tool from Princeton University to map granular GO annotations to a higher-level set of terms, thus providing a broad set of categories.



GENERIC GENE ONTOLOGY (GO) TERM MAPPER

Welcome to the GOTERMMAPPER , a tool for mapping the granular GO annotations for genes in a list to a set of broader, high-level GO parents terms (sometimes referred to as GO Slim terms), allowing you to bin y

The implementation of this Generic GO Term Mapper uses map2slim.pl script written by Chris Mungall at Berkeley Drosophila Genome Project, and some of the modules included in the GO-TermFinder distribution written by Ga GMOD project.

GO Term Mapper serves a different function than the GO Term Finder. GO Term Mapper simply bins the submitted gene list to a static set of ancestor GO terms. In contrast, GO Term Finder finds the GO terms significantly enric

Basic Inputs

1. Either Enter List of Genes (separate each gene by a return). SGD sample gene list (optional)

```
RNF213
SARS
ZNF252
TMED10P
ABCF2
```

OR Upload a File Containing List of Genes: Choose File No file chosen [CLEAR]

2. Choose 1 of the 3 Ontology Aspects: ● Process ○ Function ○ Component

3. Organism (Annotation): Saccharomyces cerevisiae (SGD)

4. Ontology: ● Generic slim ○ Yeast slim

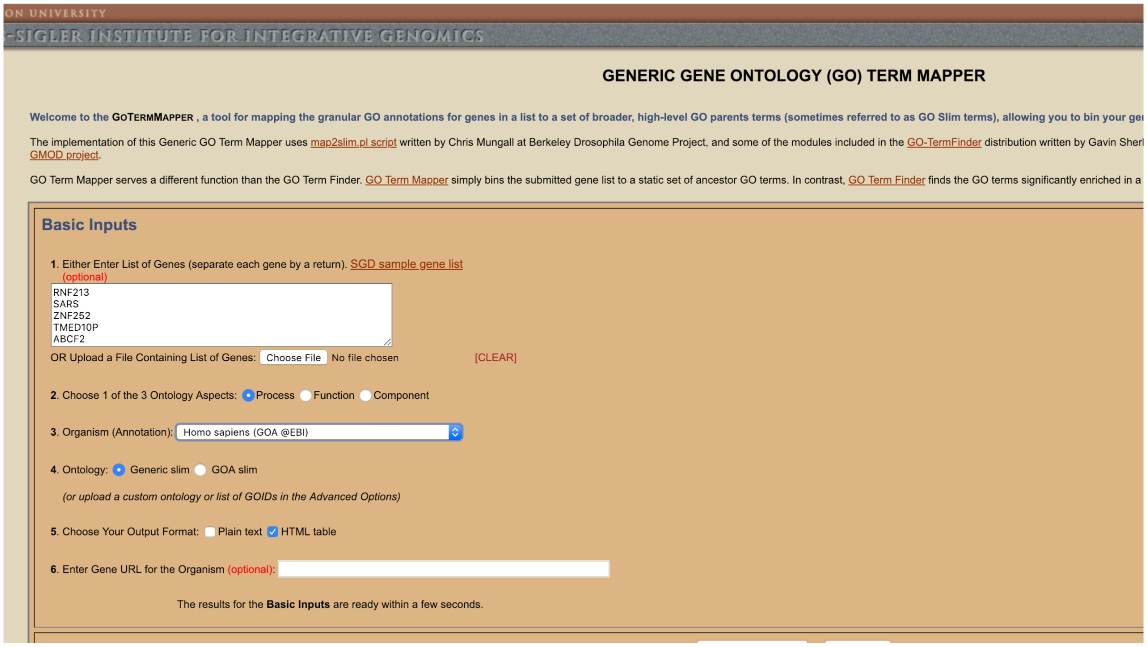
(or upload a custom ontology or list of GOIDs in the Advanced Options)

5. Choose Your Output Format: □ Plain text ☑ HTML table

6. Enter Gene URL for the Organism (optional):

The results for the **Basic Inputs** are ready within a few seconds.

Make sure to select HomoSapiens for the Organism as shown



GENERIC GENE ONTOLOGY (GO) TERM MAPPER

Welcome to the GOTERMMAPPER , a tool for mapping the granular GO annotations for genes in a list to a set of broader, high-level GO parents terms (sometimes referred to as GO Slim terms), allowing you to bin your ge

The implementation of this Generic GO Term Mapper uses map2slim.pl script written by Chris Mungall at Berkeley Drosophila Genome Project, and some of the modules included in the GO-TermFinder distribution written by Gavin Sherl GMOD project.

GO Term Mapper serves a different function than the GO Term Finder. GO Term Mapper simply bins the submitted gene list to a static set of ancestor GO terms. In contrast, GO Term Finder finds the GO terms significantly enriched in a

Basic Inputs

1. Either Enter List of Genes (separate each gene by a return). SGD sample gene list (optional)

```
RNF213
SARS
ZNF252
TMED10P
ABCF2
```

OR Upload a File Containing List of Genes: Choose File No file chosen [CLEAR]

2. Choose 1 of the 3 Ontology Aspects: ● Process ○ Function ○ Component

3. Organism (Annotation): Homo sapiens (GOA @EBI)

4. Ontology: ● Generic slim ○ GOA slim

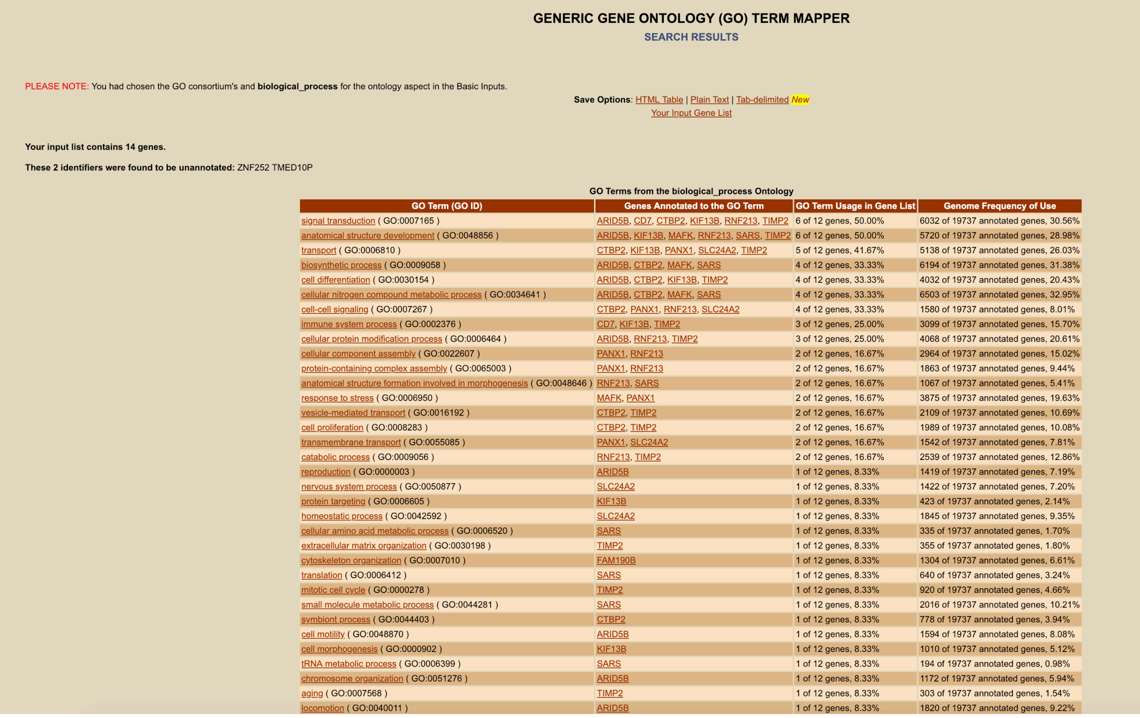
(or upload a custom ontology or list of GOIDs in the Advanced Options)

5. Choose Your Output Format: □ Plain text ☑ HTML table

6. Enter Gene URL for the Organism (optional):

The results for the **Basic Inputs** are ready within a few seconds.

The results are shown below.



GENERIC GENE ONTOLOGY (GO) TERM MAPPER

SEARCH RESULTS

PLEASE NOTE: You had chosen the GO consortium's and **biological_process** for the ontology aspect in the Basic Inputs.

Save Options: HTML Table | Plain Text | Tab-delimited *New*
Your Input Gene List

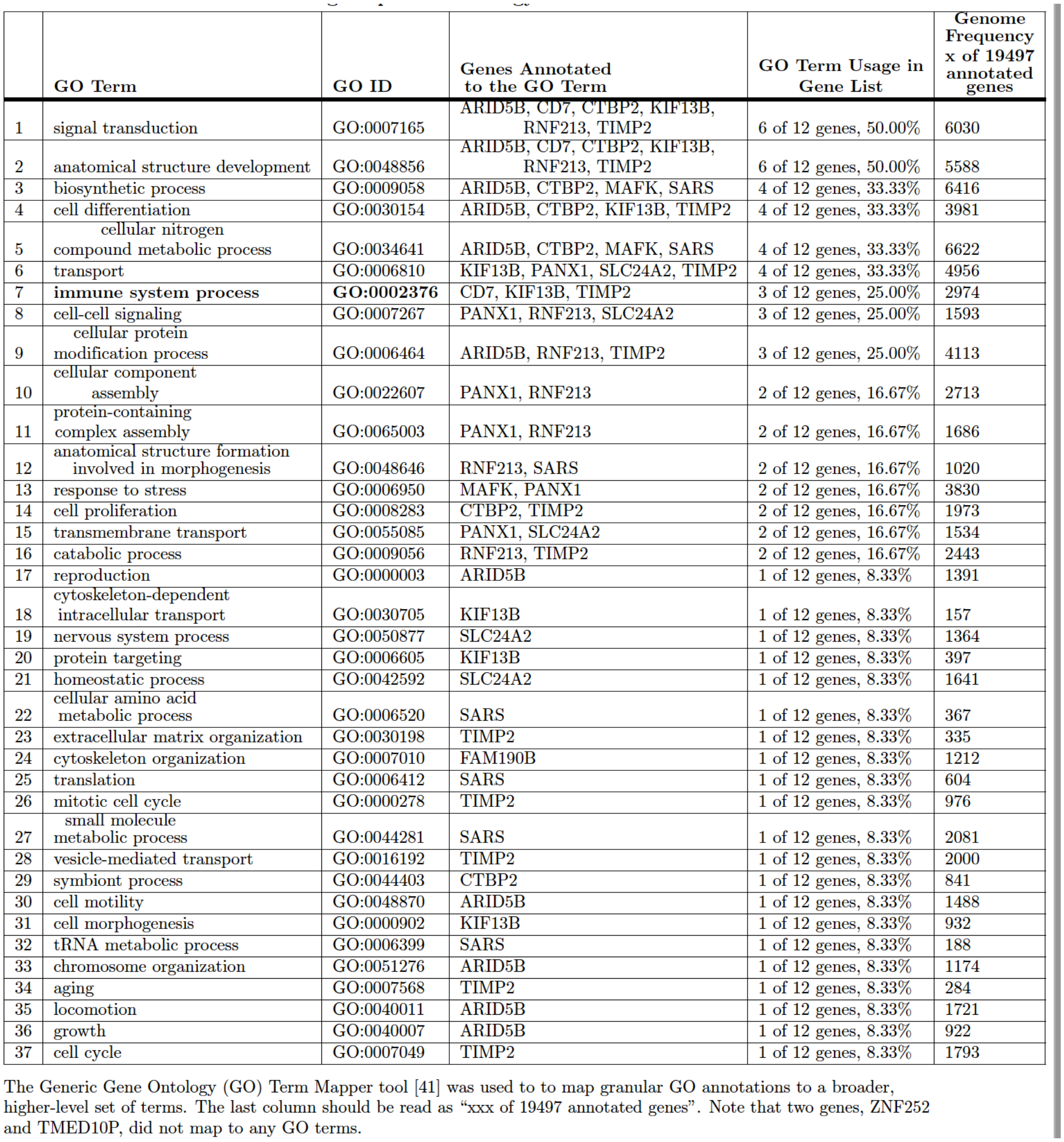
Your input list contains 14 genes.

These 2 identifiers were found to be unannotated: ZNF252 TMED10P

GO Terms from the biological_process Ontology

GO Term (GO ID)	Genes Annotated to the GO Term	GO Term Usage in Gene List	Genome Frequency of Use
signal transduction (GO:0007165)	ARID5B, CD7, CTBP2, KIF13B, RNF213, TIMP2	6 of 12 genes, 50.00%	6032 of 19737 annotated genes, 30.56%
anatomical structure development (GO:0048856)	ARID5B, KIF13B, MAFK, RNF213, SARS, TIMP2	6 of 12 genes, 50.00%	5720 of 19737 annotated genes, 28.98%
transport (GO:0006810)	CTBP2, KIF13B, PANX1, SLC24A2, TIMP2	5 of 12 genes, 41.67%	5138 of 19737 annotated genes, 26.03%
biosynthetic process (GO:0009058)	ARID5B, CTBP2, MAFK, SARS	4 of 12 genes, 33.33%	6194 of 19737 annotated genes, 31.38%
cell differentiation (GO:0030154)	ARID5B, CTBP2, KIF13B, TIMP2	4 of 12 genes, 33.33%	4032 of 19737 annotated genes, 20.43%
cellular nitrogen compound metabolic process (GO:0034641)	ARID5B, CTBP2, MAFK, SARS	4 of 12 genes, 33.33%	6503 of 19737 annotated genes, 32.95%
cell-cell signaling (GO:0007267)	CTBP2, PANX1, RNF213, SLC24A2	4 of 12 genes, 33.33%	1580 of 19737 annotated genes, 8.01%
immune system process (GO:0002376)	CD7, KIF13B, TIMP2	3 of 12 genes, 25.00%	3099 of 19737 annotated genes, 15.70%
cellular protein modification process (GO:0006464)	ARID5B, RNF213, TIMP2	3 of 12 genes, 25.00%	4068 of 19737 annotated genes, 20.61%
cellular component assembly (GO:0022607)	PANX1, RNF213	2 of 12 genes, 16.67%	2964 of 19737 annotated genes, 15.02%
protein-containing complex assembly (GO:0065003)	PANX1, RNF213	2 of 12 genes, 16.67%	1863 of 19737 annotated genes, 9.44%
anatomical structure formation involved in morphogenesis (GO:0048646)	RNF213, SARS	2 of 12 genes, 16.67%	1067 of 19737 annotated genes, 5.41%
response to stress (GO:0006950)	MAFK, PANX1	2 of 12 genes, 16.67%	3875 of 19737 annotated genes, 19.63%
vesicle-mediated transport (GO:0016192)	CTBP2, TIMP2	2 of 12 genes, 16.67%	2109 of 19737 annotated genes, 10.69%
cell proliferation (GO:0008283)	CTBP2, TIMP2	2 of 12 genes, 16.67%	1989 of 19737 annotated genes, 10.08%
transmembrane transport (GO:0055085)	PANX1, SLC24A2	2 of 12 genes, 16.67%	1542 of 19737 annotated genes, 7.81%
catabolic process (GO:0009056)	RNF213, TIMP2	2 of 12 genes, 16.67%	2539 of 19737 annotated genes, 12.86%
reproduction (GO:0000003)	ARID5B	1 of 12 genes, 8.33%	1419 of 19737 annotated genes, 7.19%
nervous system process (GO:0050877)	SLC24A2	1 of 12 genes, 8.33%	1422 of 19737 annotated genes, 7.20%
protein targeting (GO:0006605)	KIF13B	1 of 12 genes, 8.33%	423 of 19737 annotated genes, 2.14%
homeostatic process (GO:0042592)	SLC24A2	1 of 12 genes, 8.33%	1845 of 19737 annotated genes, 9.35%
cellular amino acid metabolic process (GO:0006520)	SARS	1 of 12 genes, 8.33%	335 of 19737 annotated genes, 1.70%
extracellular matrix organization (GO:0030198)	TIMP2	1 of 12 genes, 8.33%	355 of 19737 annotated genes, 1.80%
cytoskeleton organization (GO:0007010)	FAM190B	1 of 12 genes, 8.33%	1304 of 19737 annotated genes, 6.61%
translation (GO:0006412)	SARS	1 of 12 genes, 8.33%	640 of 19737 annotated genes, 3.24%
mitotic cell cycle (GO:0000278)	TIMP2	1 of 12 genes, 8.33%	920 of 19737 annotated genes, 4.66%
small molecule metabolic process (GO:0044281)	SARS	1 of 12 genes, 8.33%	2016 of 19737 annotated genes, 10.21%
symbiont process (GO:0044403)	CTBP2	1 of 12 genes, 8.33%	778 of 19737 annotated genes, 3.94%
cell motility (GO:0048870)	ARID5B	1 of 12 genes, 8.33%	1594 of 19737 annotated genes, 8.08%
cell morphogenesis (GO:0000902)	KIF13B	1 of 12 genes, 8.33%	1010 of 19737 annotated genes, 5.12%
tRNA metabolic process (GO:0006399)	SARS	1 of 12 genes, 8.33%	194 of 19737 annotated genes, 0.98%
chromosome organization (GO:0051276)	ARID5B	1 of 12 genes, 8.33%	1172 of 19737 annotated genes, 5.94%
aging (GO:0007568)	TIMP2	1 of 12 genes, 8.33%	303 of 19737 annotated genes, 1.54%
locomotion (GO:0040011)	ARID5B	1 of 12 genes, 8.33%	1820 of 19737 annotated genes, 9.22%

The results are summarized in the following table.



	GO Term	GO ID	Genes Annotated to the GO Term	GO Term Usage in Gene List	Genome Frequency x of 19497 annotated genes
1	signal transduction	GO:0007165	ARID5B, CD7, CTBP2, KIF13B, RNF213, TIMP2	6 of 12 genes, 50.00%	6030
2	anatomical structure development	GO:0048856	ARID5B, CD7, CTBP2, KIF13B, RNF213, TIMP2	6 of 12 genes, 50.00%	5588
3	biosynthetic process	GO:0009058	ARID5B, CTBP2, MAFK, SARS	4 of 12 genes, 33.33%	6416
4	cell differentiation	GO:0030154	ARID5B, CTBP2, KIF13B, TIMP2	4 of 12 genes, 33.33%	3981
5	cellular nitrogen compound metabolic process	GO:0034641	ARID5B, CTBP2, MAFK, SARS	4 of 12 genes, 33.33%	6622
6	transport	GO:0006810	KIF13B, PANX1, SLC24A2, TIMP2	4 of 12 genes, 33.33%	4956
7	**immune system process**	**GO:0002376**	CD7, KIF13B, TIMP2	3 of 12 genes, 25.00%	2974
8	cell-cell signaling	GO:0007267	PANX1, RNF213, SLC24A2	3 of 12 genes, 25.00%	1593
9	cellular protein modification process	GO:0006464	ARID5B, RNF213, TIMP2	3 of 12 genes, 25.00%	4113
10	cellular component assembly	GO:0022607	PANX1, RNF213	2 of 12 genes, 16.67%	2713
11	protein-containing complex assembly	GO:0065003	PANX1, RNF213	2 of 12 genes, 16.67%	1686
12	anatomical structure formation involved in morphogenesis	GO:0048646	RNF213, SARS	2 of 12 genes, 16.67%	1020
13	response to stress	GO:0006950	MAFK, PANX1	2 of 12 genes, 16.67%	3830
14	cell proliferation	GO:0008283	CTBP2, TIMP2	2 of 12 genes, 16.67%	1973
15	transmembrane transport	GO:0055085	PANX1, SLC24A2	2 of 12 genes, 16.67%	1534
16	catabolic process	GO:0009056	RNF213, TIMP2	2 of 12 genes, 16.67%	2443
17	reproduction	GO:0000003	ARID5B	1 of 12 genes, 8.33%	1391
18	cytoskeleton-dependent intracellular transport	GO:0030705	KIF13B	1 of 12 genes, 8.33%	157
19	nervous system process	GO:0050877	SLC24A2	1 of 12 genes, 8.33%	1364
20	protein targeting	GO:0006605	KIF13B	1 of 12 genes, 8.33%	397
21	homeostatic process	GO:0042592	SLC24A2	1 of 12 genes, 8.33%	1641
22	cellular amino acid metabolic process	GO:0006520	SARS	1 of 12 genes, 8.33%	367
23	extracellular matrix organization	GO:0030198	TIMP2	1 of 12 genes, 8.33%	335
24	cytoskeleton organization	GO:0007010	FAM190B	1 of 12 genes, 8.33%	1212
25	translation	GO:0006412	SARS	1 of 12 genes, 8.33%	604
26	mitotic cell cycle	GO:0000278	TIMP2	1 of 12 genes, 8.33%	976
27	small molecule metabolic process	GO:0044281	SARS	1 of 12 genes, 8.33%	2081
28	vesicle-mediated transport	GO:0016192	TIMP2	1 of 12 genes, 8.33%	2000
29	symbiont process	GO:0044403	CTBP2	1 of 12 genes, 8.33%	841
30	cell motility	GO:0048870	ARID5B	1 of 12 genes, 8.33%	1488
31	cell morphogenesis	GO:0000902	KIF13B	1 of 12 genes, 8.33%	932
32	tRNA metabolic process	GO:0006399	SARS	1 of 12 genes, 8.33%	188
33	chromosome organization	GO:0051276	ARID5B	1 of 12 genes, 8.33%	1174
34	aging	GO:0007568	TIMP2	1 of 12 genes, 8.33%	284
35	locomotion	GO:0040011	ARID5B	1 of 12 genes, 8.33%	1721
36	growth	GO:0040007	ARID5B	1 of 12 genes, 8.33%	922
37	cell cycle	GO:0007049	TIMP2	1 of 12 genes, 8.33%	1793

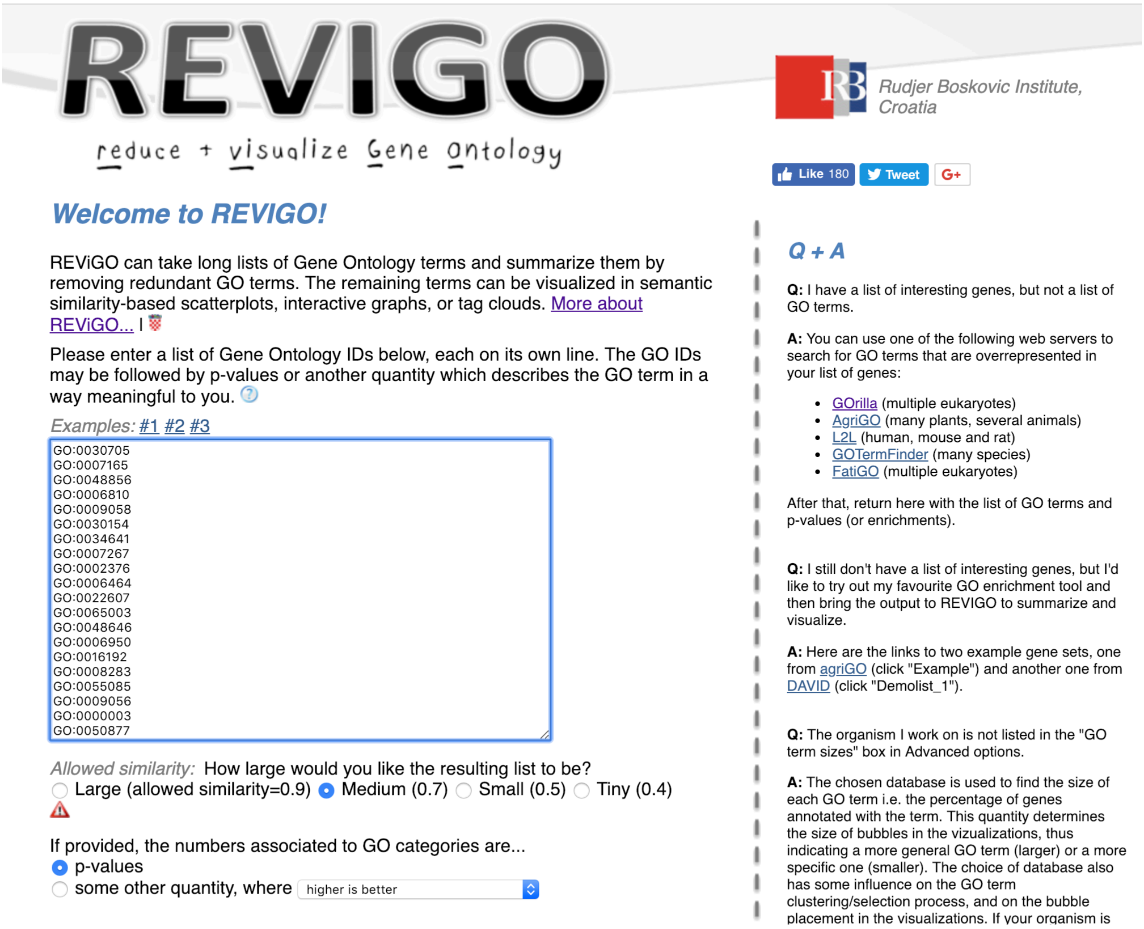
The Generic Gene Ontology (GO) Term Mapper tool [41] was used to to map granular GO annotations to a broader, higher-level set of terms. The last column should be read as "xxx of 19497 annotated genes". Note that two genes, ZNF252 and TMED10P, did not map to any GO terms.

The resulting GO terms are

GO:0030705
GO:0007165
GO:0048856
GO:0006810
GO:0009058
GO:0030154
GO:0034641
GO:0007267

GO:0002376
GO:0006464
GO:0022607
GO:0065003
GO:0048646
GO:0006950
GO:0016192
GO:0008283
GO:0055085
GO:0009056
GO:0000003
GO:0050877
GO:0006605
GO:0042592
GO:0006520
GO:0030198
GO:0007010
GO:0006412
GO:0000278
GO:0044281
GO:0044403
GO:0048870
GO:0000902
GO:0006399
GO:0051276
GO:0007568
GO:0040011
GO:0040007
GO:0007049

Next, the **REVIGO**, an online tool that summarizes and visualizes lists of gene ontology terms, to find a representative set of terms using a clustering algorithm.



REVIGO
reduce + visualize Gene ontology
Rudjer Boskovic Institute, Croatia
Like 180
Tweet
G+

Welcome to REVIGO!

REViGO can take long lists of Gene Ontology terms and summarize them by removing redundant GO terms. The remaining terms can be visualized in semantic similarity-based scatterplots, interactive graphs, or tag clouds. More about REViGO... |

Please enter a list of Gene Ontology IDs below, each on its own line. The GO IDs may be followed by p-values or another quantity which describes the GO term in a way meaningful to you.

Examples: #1 #2 #3

```
GO:0030705
GO:0007165
GO:0048856
GO:0006810
GO:0009058
GO:0030154
GO:0034641
GO:0007267
GO:0002376
GO:0006464
GO:0022607
GO:0065003
GO:0048646
GO:0006950
GO:0016192
GO:0008283
GO:0055085
GO:0009056
GO:0000003
GO:0050877
```

Allowed similarity: How large would you like the resulting list to be?
○ Large (allowed similarity=0.9) ● Medium (0.7) ○ Small (0.5) ○ Tiny (0.4)

If provided, the numbers associated to GO categories are...
● p-values
○ some other quantity, where higher is better

Q + A

Q: I have a list of interesting genes, but not a list of GO terms.

A: You can use one of the following web servers to search for GO terms that are overrepresented in your list of genes:

- GOrilla (multiple eukaryotes)
- AgriGO (many plants, several animals)
- L2L (human, mouse and rat)
- GOTermFinder (many species)
- FatiGO (multiple eukaryotes)

After that, return here with the list of GO terms and p-values (or enrichments).

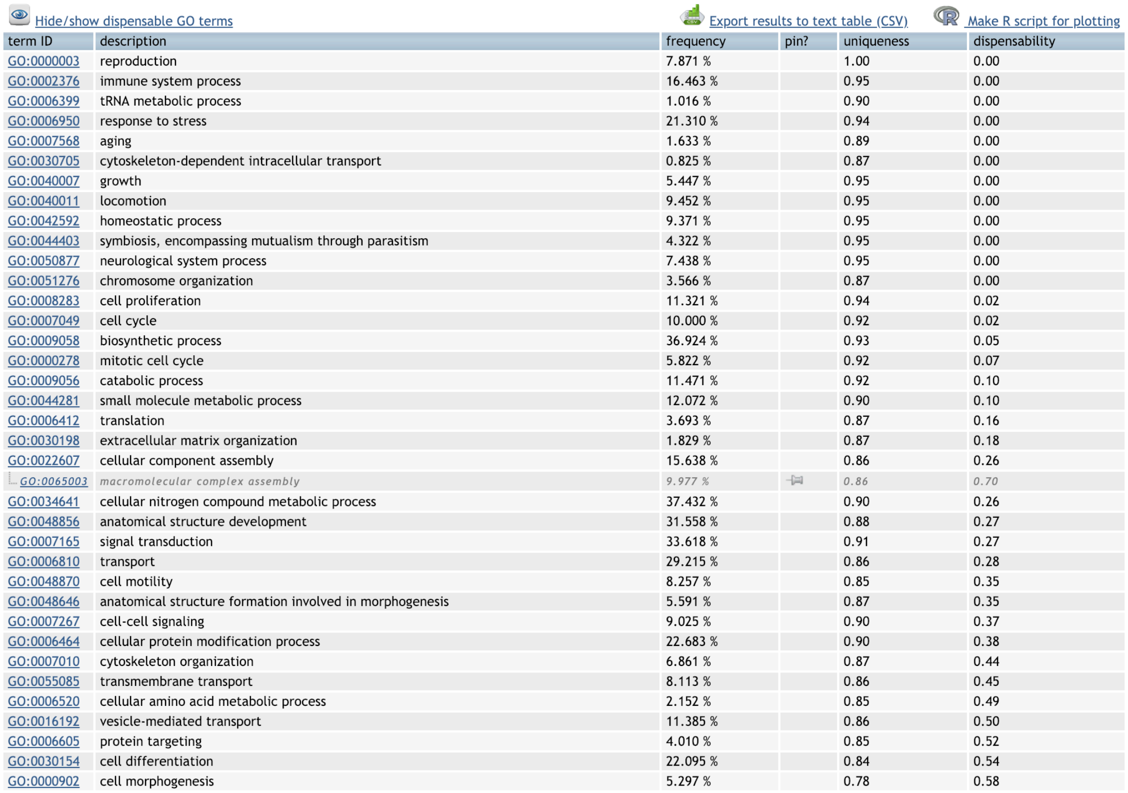
Q: I still don't have a list of interesting genes, but I'd like to try out my favourite GO enrichment tool and then bring the output to REVIGO to summarize and visualize.

A: Here are the links to two example gene sets, one from agriGO (click "Example") and another one from DAVID (click "Demolist_1").

Q: The organism I work on is not listed in the "GO term sizes" box in Advanced options.

A: The chosen database is used to find the size of each GO term i.e. the percentage of genes annotated with the term. This quantity determines the size of bubbles in the vizualizations, thus indicating a more general GO term (larger) or a more specific one (smaller). The choice of database also has some influence on the GO term clustering/selection process, and on the bubble placement in the visualizations. If your organism is

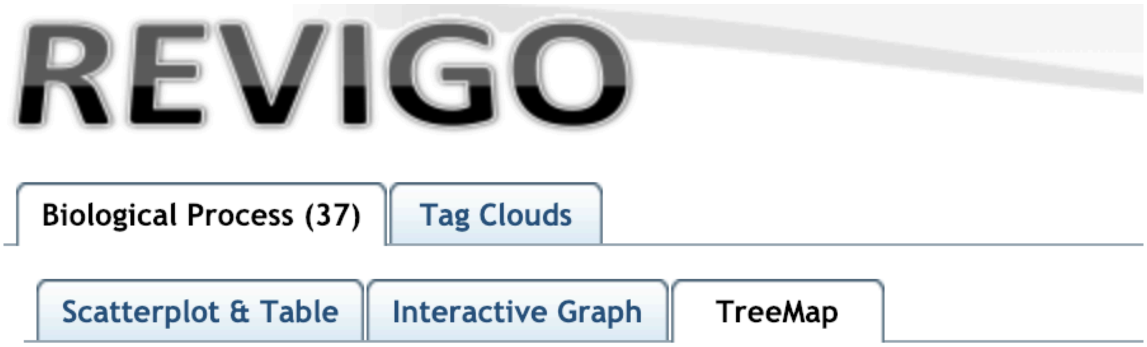
The frequency and uniqueness of the GO terms can be seen below.



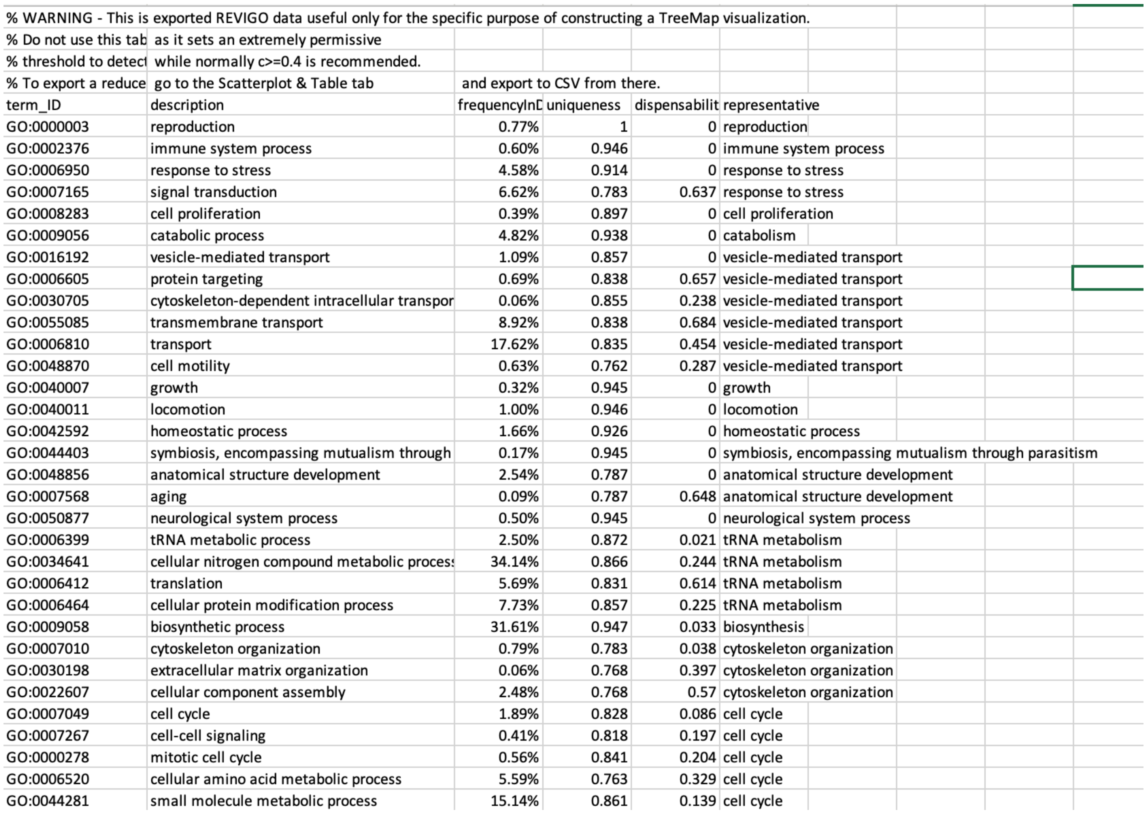
Hide/show dispensable GO terms Export results to text table (CSV) Make R script for plotting

term ID	description	frequency	pin?	uniqueness	dispensability
GO:0000003	reproduction	7.871 %		1.00	0.00
GO:0002376	immune system process	16.463 %		0.95	0.00
GO:0006399	tRNA metabolic process	1.016 %		0.90	0.00
GO:0006950	response to stress	21.310 %		0.94	0.00
GO:0007568	aging	1.633 %		0.89	0.00
GO:0030705	cytoskeleton-dependent intracellular transport	0.825 %		0.87	0.00
GO:0040007	growth	5.447 %		0.95	0.00
GO:0040011	locomotion	9.452 %		0.95	0.00
GO:0042592	homeostatic process	9.371 %		0.95	0.00
GO:0044403	symbiosis, encompassing mutualism through parasitism	4.322 %		0.95	0.00
GO:0050877	neurological system process	7.438 %		0.95	0.00
GO:0051276	chromosome organization	3.566 %		0.87	0.00
GO:0008283	cell proliferation	11.321 %		0.94	0.02
GO:0007049	cell cycle	10.000 %		0.92	0.02
GO:0009058	biosynthetic process	36.924 %		0.93	0.05
GO:0000278	mitotic cell cycle	5.822 %		0.92	0.07
GO:0009056	catabolic process	11.471 %		0.92	0.10
GO:0044281	small molecule metabolic process	12.072 %		0.90	0.10
GO:0006412	translation	3.693 %		0.87	0.16
GO:0030198	extracellular matrix organization	1.829 %		0.87	0.18
GO:0022607	cellular component assembly	15.638 %		0.86	0.26
GO:0065003	*macromolecular complex assembly*	*9.977 %*	📌	*0.86*	*0.70*
GO:0034641	cellular nitrogen compound metabolic process	37.432 %		0.90	0.26
GO:0048856	anatomical structure development	31.558 %		0.88	0.27
GO:0007165	signal transduction	33.618 %		0.91	0.27
GO:0006810	transport	29.215 %		0.86	0.28
GO:0048870	cell motility	8.257 %		0.85	0.35
GO:0048646	anatomical structure formation involved in morphogenesis	5.591 %		0.87	0.35
GO:0007267	cell-cell signaling	9.025 %		0.90	0.37
GO:0006464	cellular protein modification process	22.683 %		0.90	0.38
GO:0007010	cytoskeleton organization	6.861 %		0.87	0.44
GO:0055085	transmembrane transport	8.113 %		0.86	0.45
GO:0006520	cellular amino acid metabolic process	2.152 %		0.85	0.49
GO:0016192	vesicle-mediated transport	11.385 %		0.86	0.50
GO:0006605	protein targeting	4.010 %		0.85	0.52
GO:0030154	cell differentiation	22.095 %		0.84	0.54
GO:0000902	cell morphogenesis	5.297 %		0.78	0.58

Click on the TreeMap tab.

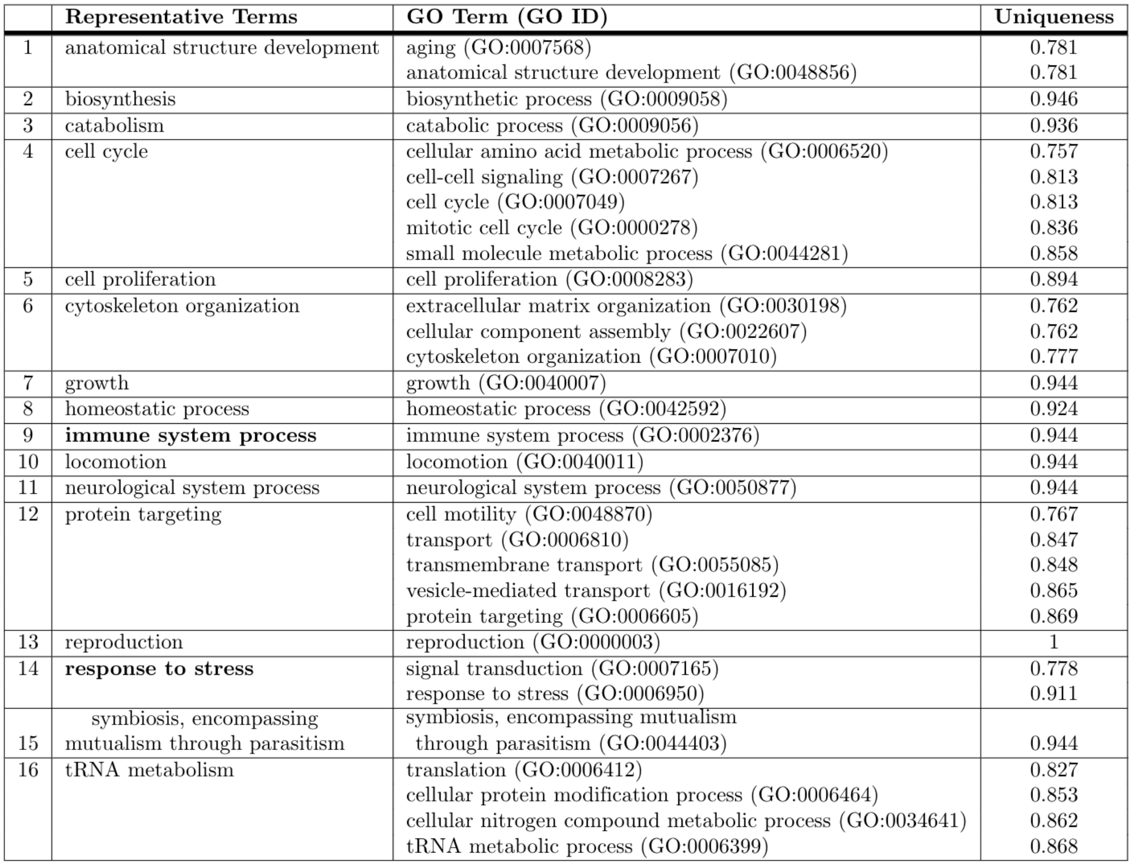


Download the csv file at the bottom of the page. You should see a csv file with the following content.



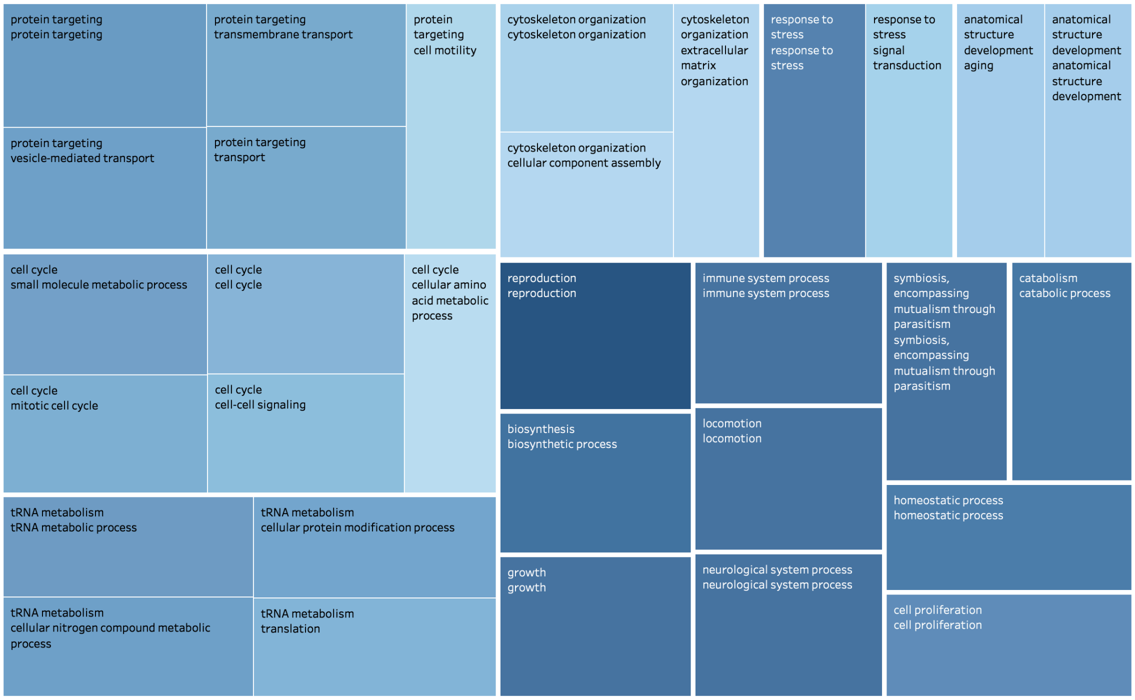
% WARNING - This is	exported REVIGO data useful only for the specific purpose of constructing a TreeMap visualization.				
% Do not use this tab	as it sets an extremely permissive				
% threshold to detect	while normally c>=0.4 is recommended.				
% To export a reduce	go to the Scatterplot & Table tab	and export to CSV from there.			
term_ID	description	frequencyInD	uniqueness	dispensabilit	representative
GO:0000003	reproduction	0.77%	1	0	reproduction
GO:0002376	immune system process	0.60%	0.946	0	immune system process
GO:0006950	response to stress	4.58%	0.914	0	response to stress
GO:0007165	signal transduction	6.62%	0.783	0.637	response to stress
GO:0008283	cell proliferation	0.39%	0.897	0	cell proliferation
GO:0009056	catabolic process	4.82%	0.938	0	catabolism
GO:0016192	vesicle-mediated transport	1.09%	0.857	0	vesicle-mediated transport
GO:0006605	protein targeting	0.69%	0.838	0.657	vesicle-mediated transport
GO:0030705	cytoskeleton-dependent intracellular transpor	0.06%	0.855	0.238	vesicle-mediated transport
GO:0055085	transmembrane transport	8.92%	0.838	0.684	vesicle-mediated transport
GO:0006810	transport	17.62%	0.835	0.454	vesicle-mediated transport
GO:0048870	cell motility	0.63%	0.762	0.287	vesicle-mediated transport
GO:0040007	growth	0.32%	0.945	0	growth
GO:0040011	locomotion	1.00%	0.946	0	locomotion
GO:0042592	homeostatic process	1.66%	0.926	0	homeostatic process
GO:0044403	symbiosis, encompassing mutualism through	0.17%	0.945	0	symbiosis, encompassing mutualism through parasitism
GO:0048856	anatomical structure development	2.54%	0.787	0	anatomical structure development
GO:0007568	aging	0.09%	0.787	0.648	anatomical structure development
GO:0050877	neurological system process	0.50%	0.945	0	neurological system process
GO:0006399	tRNA metabolic process	2.50%	0.872	0.021	tRNA metabolism
GO:0034641	cellular nitrogen compound metabolic proces	34.14%	0.866	0.244	tRNA metabolism
GO:0006412	translation	5.69%	0.831	0.614	tRNA metabolism
GO:0006464	cellular protein modification process	7.73%	0.857	0.225	tRNA metabolism
GO:0009058	biosynthetic process	31.61%	0.947	0.033	biosynthesis
GO:0007010	cytoskeleton organization	0.79%	0.783	0.038	cytoskeleton organization
GO:0030198	extracellular matrix organization	0.06%	0.768	0.397	cytoskeleton organization
GO:0022607	cellular component assembly	2.48%	0.768	0.57	cytoskeleton organization
GO:0007049	cell cycle	1.89%	0.828	0.086	cell cycle
GO:0007267	cell-cell signaling	0.41%	0.818	0.197	cell cycle
GO:0000278	mitotic cell cycle	0.56%	0.841	0.204	cell cycle
GO:0006520	cellular amino acid metabolic process	5.59%	0.763	0.329	cell cycle
GO:0044281	small molecule metabolic process	15.14%	0.861	0.139	cell cycle

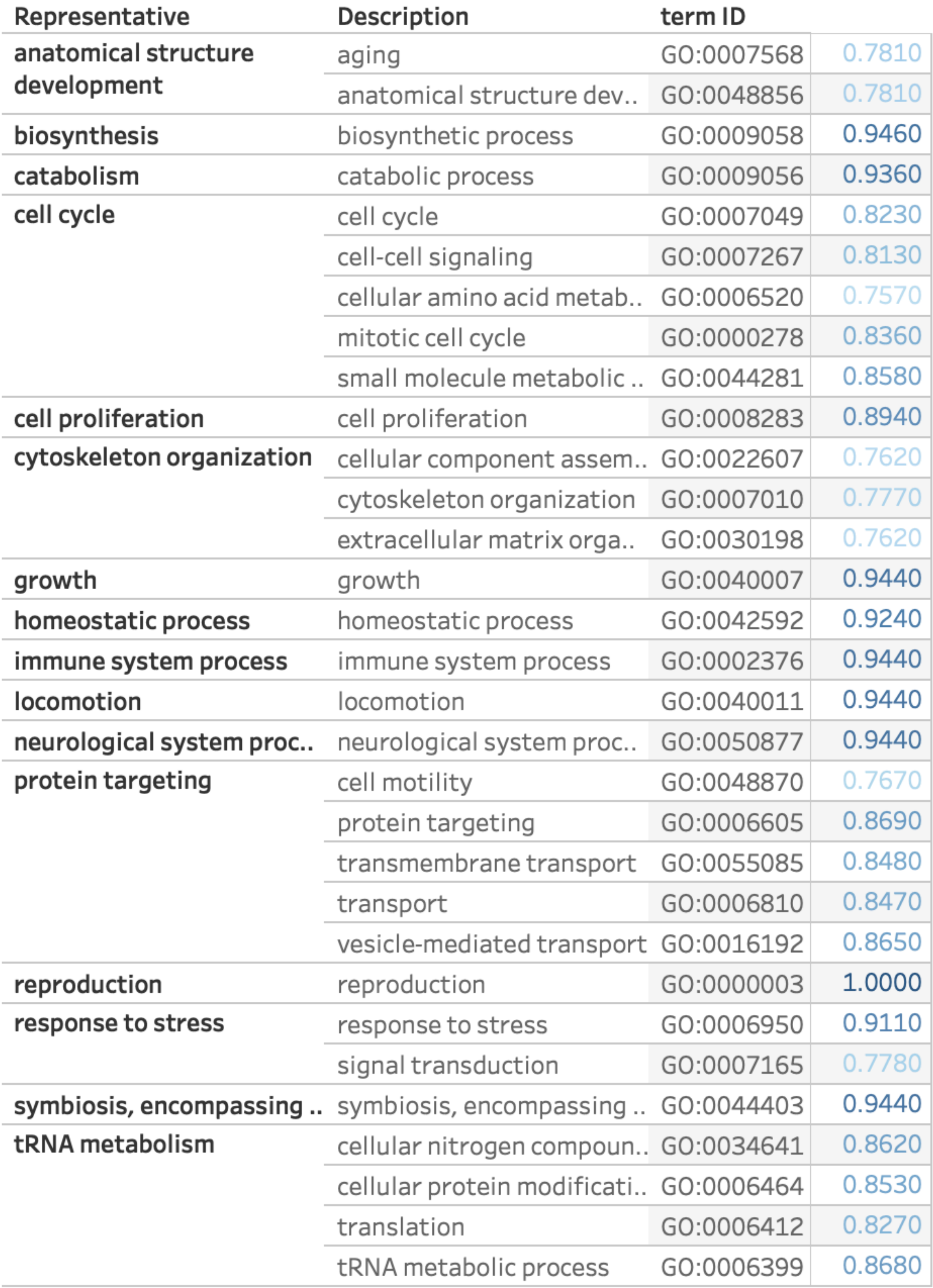
This can be summarized by the following table.



	Representative Terms	**GO Term (GO ID)**	**Uniqueness**
1	anatomical structure development	aging (GO:0007568) anatomical structure development (GO:0048856)	0.781 0.781
2	biosynthesis	biosynthetic process (GO:0009058)	0.946
3	catabolism	catabolic process (GO:0009056)	0.936
4	cell cycle	cellular amino acid metabolic process (GO:0006520) cell-cell signaling (GO:0007267) cell cycle (GO:0007049) mitotic cell cycle (GO:0000278) small molecule metabolic process (GO:0044281)	0.757 0.813 0.813 0.836 0.858
5	cell proliferation	cell proliferation (GO:0008283)	0.894
6	cytoskeleton organization	extracellular matrix organization (GO:0030198) cellular component assembly (GO:0022607) cytoskeleton organization (GO:0007010)	0.762 0.762 0.777
7	growth	growth (GO:0040007)	0.944
8	homeostatic process	homeostatic process (GO:0042592)	0.924
9	**immune system process**	immune system process (GO:0002376)	0.944
10	locomotion	locomotion (GO:0040011)	0.944
11	neurological system process	neurological system process (GO:0050877)	0.944
12	protein targeting	cell motility (GO:0048870) transport (GO:0006810) transmembrane transport (GO:0055085) vesicle-mediated transport (GO:0016192) protein targeting (GO:0006605)	0.767 0.847 0.848 0.865 0.869
13	reproduction	reproduction (GO:0000003)	1
14	**response to stress**	signal transduction (GO:0007165) response to stress (GO:0006950)	0.778 0.911
15	symbiosis, encompassing mutualism through parasitism	symbiosis, encompassing mutualism through parasitism (GO:0044403)	0.944
16	tRNA metabolism	translation (GO:0006412) cellular protein modification process (GO:0006464) cellular nitrogen compound metabolic process (GO:0034641) tRNA metabolic process (GO:0006399)	0.827 0.853 0.862 0.868

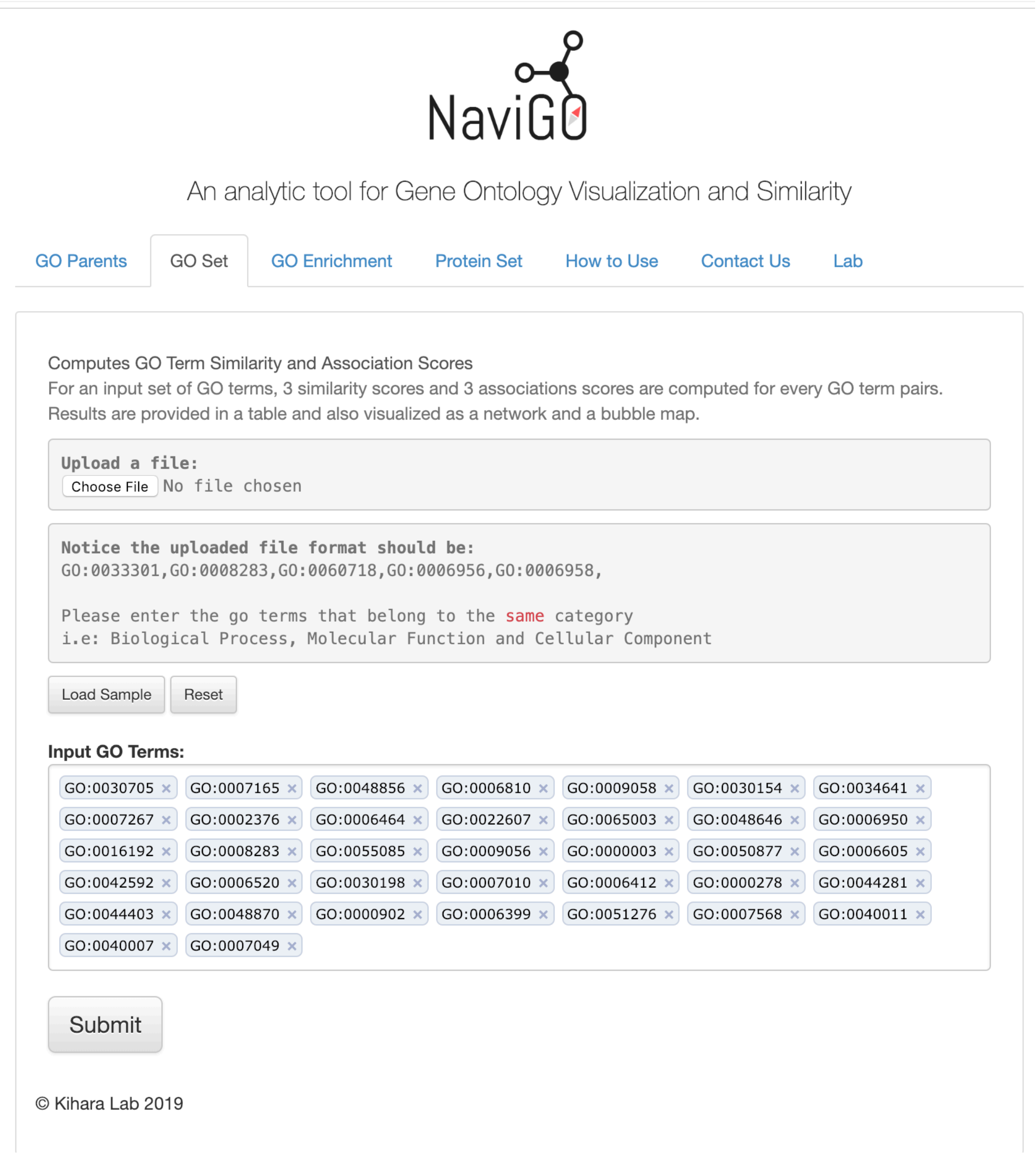
This same data can be represented in a treemap.


protein targeting
protein targeting
protein targeting
transmembrane transport
protein targeting
cell motility
protein targeting
vesicle-mediated transport
protein targeting
transport
cytoskeleton organization
cytoskeleton organization
cytoskeleton organization
extracellular matrix organization
cytoskeleton organization
cellular component assembly
response to stress
response to stress
response to stress
signal transduction
anatomical structure development
aging
anatomical structure development
anatomical structure development
cell cycle
small molecule metabolic process
cell cycle
cell cycle
cell cycle
cellular amino acid metabolic process
cell cycle
mitotic cell cycle
cell cycle
cell-cell signaling
tRNA metabolism
tRNA metabolic process
tRNA metabolism
cellular protein modification process
tRNA metabolism
cellular nitrogen compound metabolic process
tRNA metabolism
translation
reproduction
reproduction
immune system process
immune system process
symbiosis, encompassing mutualism through parasitism
symbiosis, encompassing mutualism through parasitism
catabolism
catabolic process
biosynthesis
biosynthetic process
locomotion
locomotion
homeostatic process
homeostatic process
growth
growth
neurological system process
neurological system process
cell proliferation
cell proliferation



Representative	Description	term ID	
anatomical structure development	aging	GO:0007568	0.7810
	anatomical structure dev..	GO:0048856	0.7810
biosynthesis	biosynthetic process	GO:0009058	0.9460
catabolism	catabolic process	GO:0009056	0.9360
cell cycle	cell cycle	GO:0007049	0.8230
	cell-cell signaling	GO:0007267	0.8130
	cellular amino acid metab..	GO:0006520	0.7570
	mitotic cell cycle	GO:0000278	0.8360
	small molecule metabolic ..	GO:0044281	0.8580
cell proliferation	cell proliferation	GO:0008283	0.8940
cytoskeleton organization	cellular component assem..	GO:0022607	0.7620
	cytoskeleton organization	GO:0007010	0.7770
	extracellular matrix orga..	GO:0030198	0.7620
growth	growth	GO:0040007	0.9440
homeostatic process	homeostatic process	GO:0042592	0.9240
immune system process	immune system process	GO:0002376	0.9440
locomotion	locomotion	GO:0040011	0.9440
neurological system proc..	neurological system proc..	GO:0050877	0.9440
protein targeting	cell motility	GO:0048870	0.7670
	protein targeting	GO:0006605	0.8690
	transmembrane transport	GO:0055085	0.8480
	transport	GO:0006810	0.8470
	vesicle-mediated transport	GO:0016192	0.8650
reproduction	reproduction	GO:0000003	1.0000
response to stress	response to stress	GO:0006950	0.9110
	signal transduction	GO:0007165	0.7780
symbiosis, encompassing ..	symbiosis, encompassing ..	GO:0044403	0.9440
tRNA metabolism	cellular nitrogen compoun..	GO:0034641	0.8620
	cellular protein modificati..	GO:0006464	0.8530
	translation	GO:0006412	0.8270
	tRNA metabolic process	GO:0006399	0.8680

The GO terms can be visualized using **NaviGo**.



An analytic tool for Gene Ontology Visualization and Similarity

Computes GO Term Similarity and Association Scores

For an input set of GO terms, 3 similarity scores and 3 associations scores are computed for every GO term pairs. Results are provided in a table and also visualized as a network and a bubble map.

```
Upload a file:
No file chosen
```

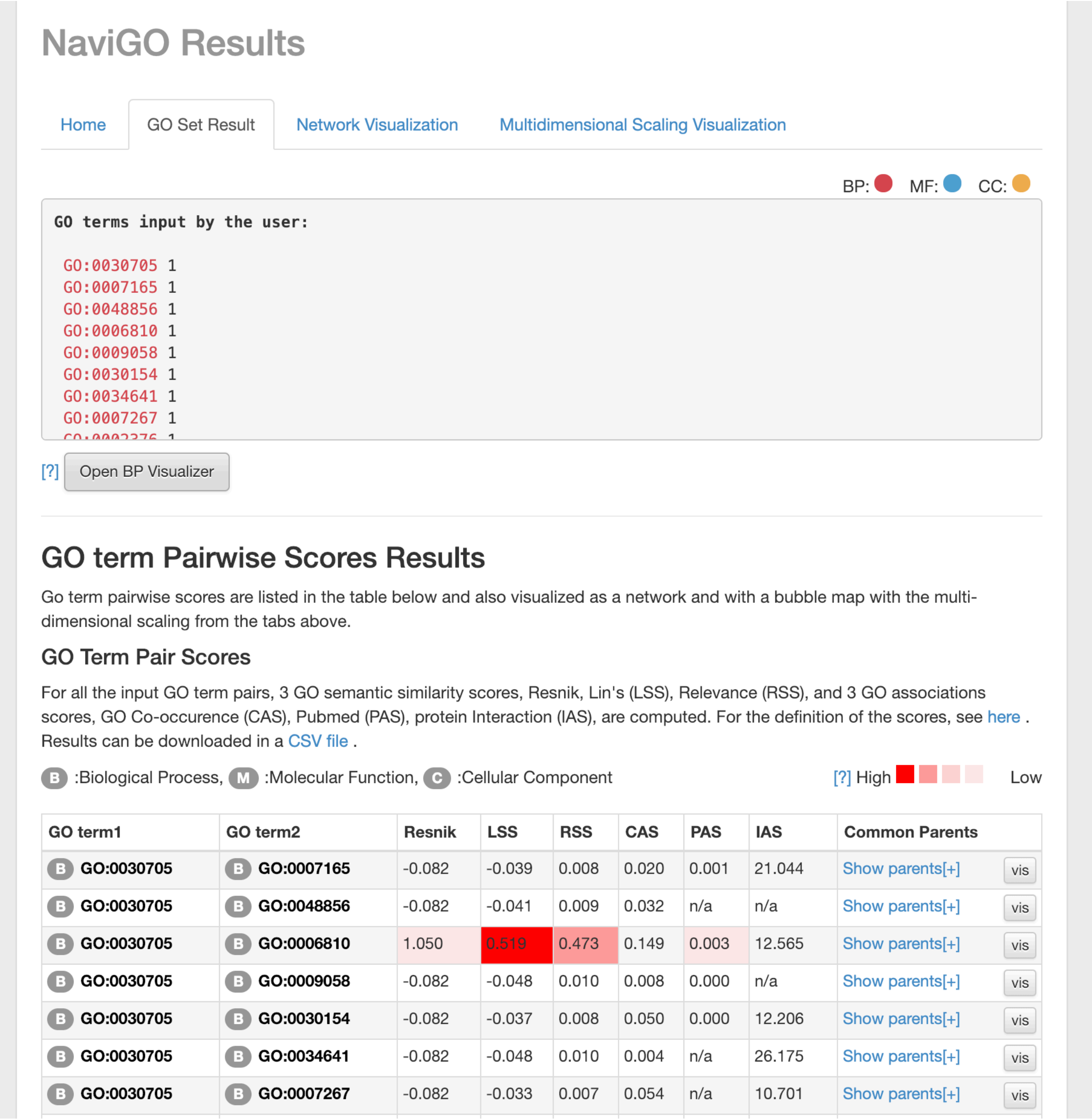
```
Notice the uploaded file format should be:
GO:0033301,GO:0008283,GO:0060718,GO:0006956,GO:0006958,

Please enter the go terms that belong to the same category
i.e: Biological Process, Molecular Function and Cellular Component
```

Input GO Terms:

GO:0030705 × GO:0007165 × GO:0048856 × GO:0006810 × GO:0009058 × GO:0030154 × GO:0034641 ×
GO:0007267 × GO:0002376 × GO:0006464 × GO:0022607 × GO:0065003 × GO:0048646 × GO:0006950 ×
GO:0016192 × GO:0008283 × GO:0055085 × GO:0009056 × GO:0000003 × GO:0050877 × GO:0006605 ×
GO:0042592 × GO:0006520 × GO:0030198 × GO:0007010 × GO:0006412 × GO:0000278 × GO:0044281 ×
GO:0044403 × GO:0048870 × GO:0000902 × GO:0006399 × GO:0051276 × GO:0007568 × GO:0040011 ×
GO:0040007 × GO:0007049 ×

© Kihara Lab 2019

Results are shown below.



NaviGO Results

```
GO terms input by the user:

GO:0030705 1
GO:0007165 1
GO:0048856 1
GO:0006810 1
GO:0009058 1
GO:0030154 1
GO:0034641 1
GO:0007267 1
GO:0002376 1
```

GO term Pairwise Scores Results

Go term pairwise scores are listed in the table below and also visualized as a network and with a bubble map with the multi-dimensional scaling from the tabs above.

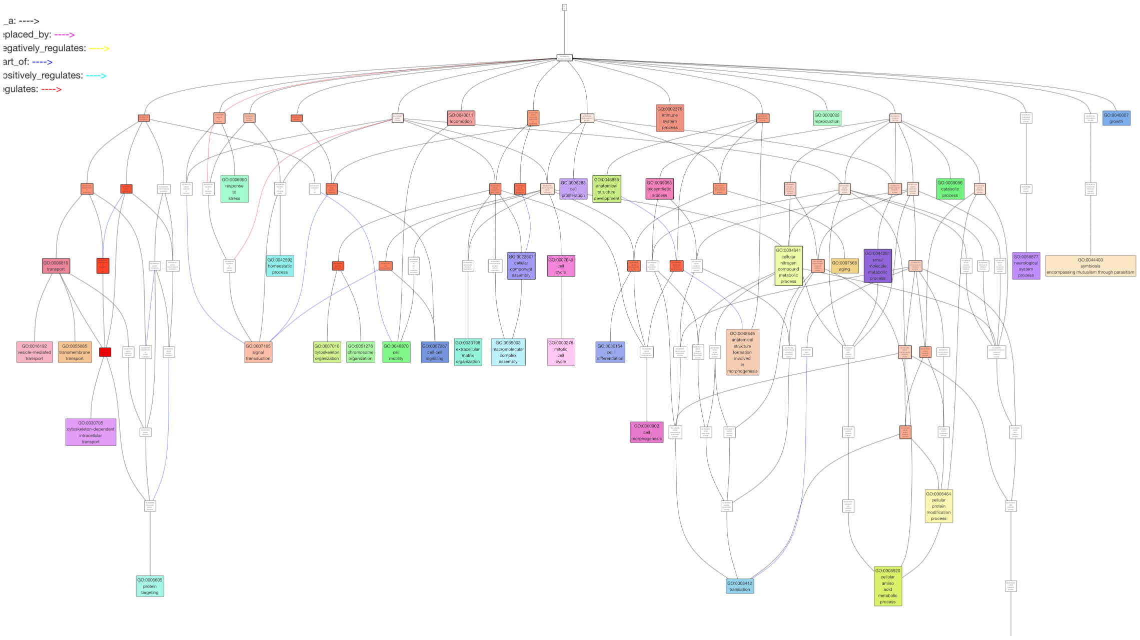
GO Term Pair Scores

For all the input GO term pairs, 3 GO semantic similarity scores, Resnik, Lin's (LSS), Relevance (RSS), and 3 GO associations scores, GO Co-occurence (CAS), Pubmed (PAS), protein Interaction (IAS), are computed. For the definition of the scores, see here . Results can be downloaded in a CSV file .

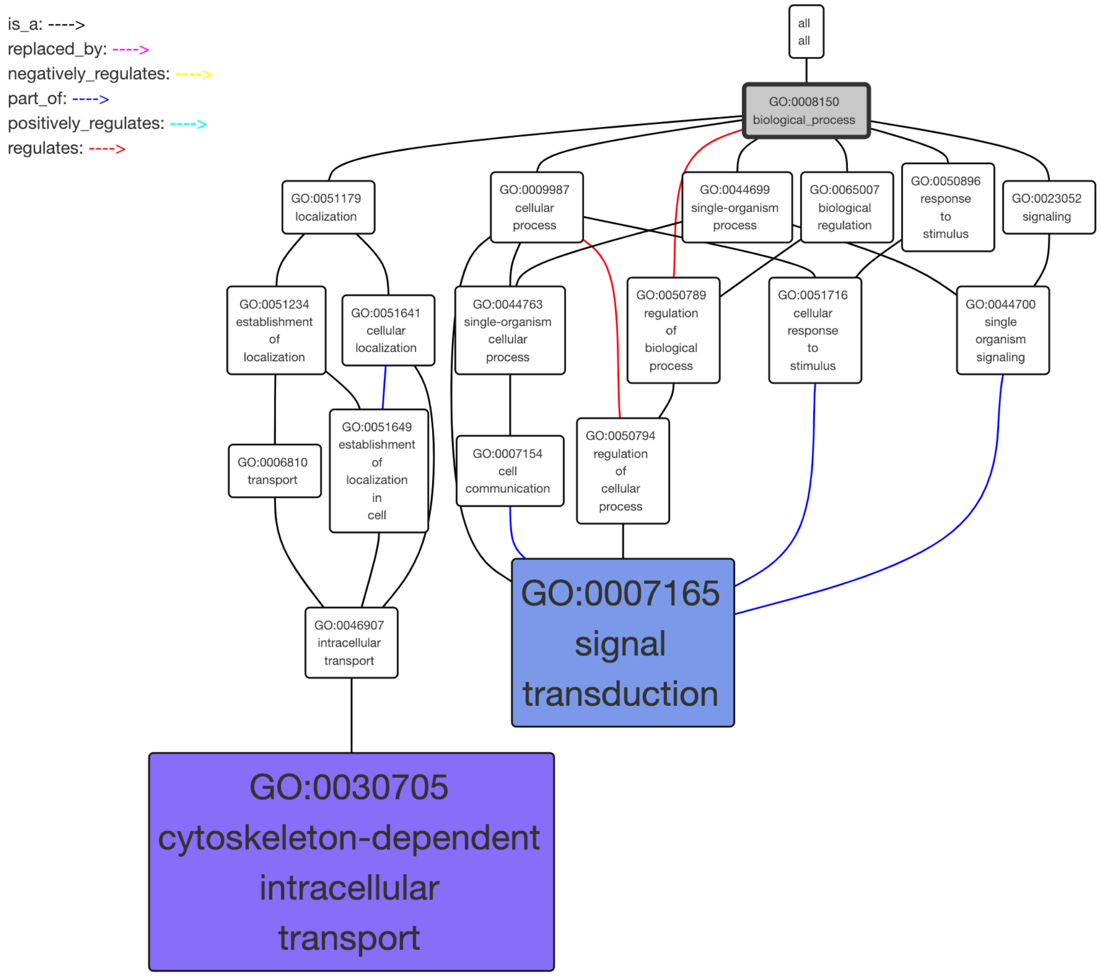
B :Biological Process, M :Molecular Function, C :Cellular Component [?] High 🟥 🟥 🟥 🟥 Low

GO term1	GO term2	Resnik	LSS	RSS	CAS	PAS	IAS	Common Parents	
B GO:0030705	B GO:0007165	-0.082	-0.039	0.008	0.020	0.001	21.044	Show parents[+]	vis
B GO:0030705	B GO:0048856	-0.082	-0.041	0.009	0.032	n/a	n/a	Show parents[+]	vis
B GO:0030705	B GO:0006810	1.050	0.519	0.473	0.149	0.003	12.565	Show parents[+]	vis
B GO:0030705	B GO:0009058	-0.082	-0.048	0.010	0.008	0.000	n/a	Show parents[+]	vis
B GO:0030705	B GO:0030154	-0.082	-0.037	0.008	0.050	0.000	12.206	Show parents[+]	vis
B GO:0030705	B GO:0034641	-0.082	-0.048	0.010	0.004	n/a	26.175	Show parents[+]	vis
B GO:0030705	B GO:0007267	-0.082	-0.033	0.007	0.054	n/a	10.701	Show parents[+]	vis

Click on Open BP Visualizer to see the network of terms.


_a: ---->
eplaced_by: ---->
egatively_regulates: ---->
art_of: ---->
ositively_regulates: ---->
egulates: ---->
GO:0040011 locomotion
GO:0002376 immune system process
GO:0000003 reproduction
GO:0040007 growth
GO:0006950 response to stress
GO:0008283 cell proliferation
GO:0048856 anatomical structure development
GO:0009058 biosynthetic process
GO:0009056 catabolic process
GO:0006810 transport
GO:0042592 homeostatic process
GO:0022607 cellular component assembly
GO:0007049 cell cycle
GO:0034641 cellular nitrogen compound metabolic process
GO:0007568 aging
GO:0044281 small molecule metabolic process
GO:0050877 neurological system process
GO:0044403 symbiosis encompassing mutualism through parasitism
GO:0016192 vesicle-mediated transport
GO:0055085 transmembrane transport
GO:0007165 signal transduction
GO:0007010 cytoskeleton organization
GO:0051276 chromosome organization
GO:0048870 cell motility
GO:0007267 cell-cell signaling
GO:0030198 extracellular matrix organization
GO:0065003 macromolecular complex assembly
GO:0000278 mitotic cell cycle
GO:0030154 cell differentiation
GO:0048646 anatomical structure formation involved in morphogenesis
GO:0030705 cytoskeleton-dependent intracellular transport
GO:0000902 cell morphogenesis
GO:0006464 cellular protein modification process
GO:0006605 protein targeting
GO:0006412 translation
GO:0006520 cellular amino acid metabolic process

Single GO terms can also be visualized, as shown below.


is_a: ---->
replaced_by: ---->
negatively_regulates: ---->
part_of: ---->
positively_regulates: ---->
regulates: ---->
all all
GO:0008150 biological_process
GO:0051179 localization
GO:0009987 cellular process
GO:0044699 single-organism process
GO:0065007 biological regulation
GO:0050896 response to stimulus
GO:0023052 signaling
GO:0051234 establishment of localization
GO:0051641 cellular localization
GO:0044763 single-organism cellular process
GO:0050789 regulation of biological process
GO:0051716 cellular response to stimulus
GO:0044700 single organism signaling
GO:0006810 transport
GO:0051649 establishment of localization in cell
GO:0007154 cell communication
GO:0050794 regulation of cellular process
GO:0007165 signal transduction
GO:0046907 intracellular transport
GO:0030705 cytoskeleton-dependent intracellular transport

See https://link.springer.com/protocol/10.1007/978-1-4939-3743-1_15 for a list of GO visualization tools.