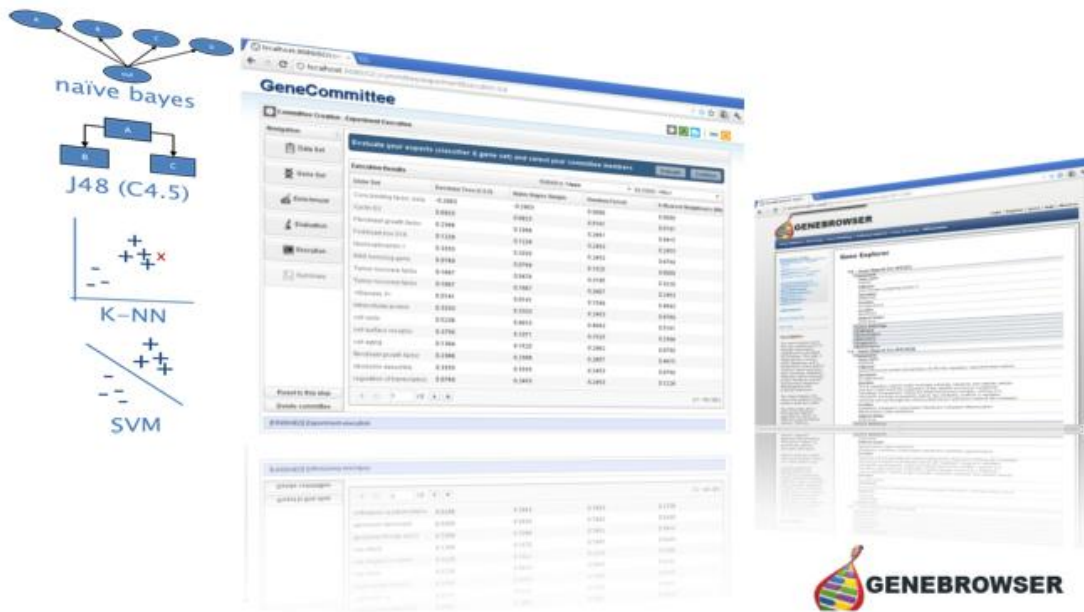


# GENECOMMITTEE



<http://sing.ei.uvigo.es/GC/>

## Expert Manual

*This document will guide you through a step-by-step tutorial showing the capabilities of GENECOMMITTEE for extensively testing the discriminatory power of biologically relevant gene sets in microarray data classification.*

# Contents

About the tool .....	1
Login.....	2
Welcome to GENECOMMITTEE .....	3
Committee Training.....	4
Data Set Selection.....	4
Gene Selection.....	5
Enriched Gene Set .....	6
Evaluation Configuration .....	8
Experiment Execution .....	9
Committee Summary.....	10
Diagnostic Mode .....	12
Data Management.....	15
Data Set Format .....	16

## About the tool

GENECOMMITTEE is a web-based interactive tool giving specific support to study of the discriminative classification power of custom hypothesis in the form of biologically relevant gene sets. With a straightforward and intuitive interface, GENECOMMITTEE is able to provide valuable information for diagnostic analyses and clinical management decisions based on systematically evaluating custom hypothesis over different data sets using complementary classifiers, a key aspect in clinical research.

### **GENECOMMITTEE's main features:**

- *Upload, store and manage different microarray data files.*
- *Several configurable classification techniques, including Naïve Bayes, Decision Trees (C4.5/J48), K-NN (K-nearest neighbours) and SVM (Support Vector Machines).*
- *User-friendly 6-step wizard to create new committees.*
- *Classifier evaluations are executed in parallel in our server.*
- *Email notifications, containing a direct link to the results stored in the server.*

### **GeneBrowser integration:**

GENECOMMITTEE application is interconnected with the successful GeneBrowser server, a web-based tool for gene set enrichment.

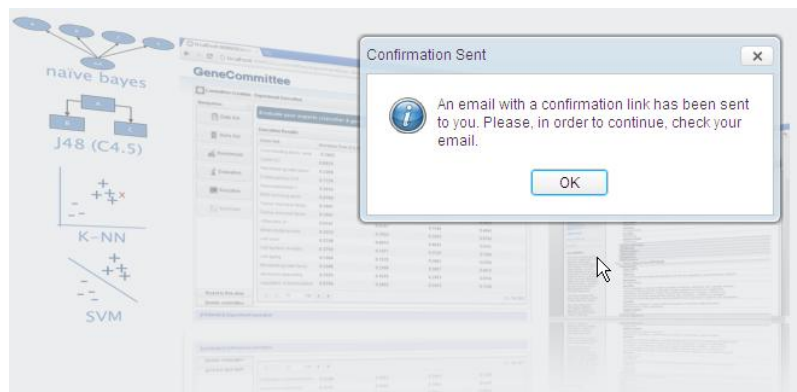
## Login

On the welcome page the user can choose to either login using a demo account or create a new account. Using the demo account (the fields are already fulfilled), the user will have access to several data sets and committees previously created.

To create a new account the login form available on the home page can be used. The account information must be based on the user's email address, which will be used to send login information, and processing results, when completed. Each registered user will have an independent workspace inside the GENECOMMITTEE server.

The screenshot shows a web interface for logging in. It has two main sections. The top section is titled 'Guest Login' and contains a blue button that says 'Try GeneCommittee!' and a grey button that says 'Enter as guest'. The bottom section is titled 'Login' and contains two input fields: 'Email' and 'Password'. Below the 'Password' field is a grey button labeled 'Login'. At the bottom of the 'Login' section are two blue links: 'Forgot my password' and 'New Account'.

An email message is sent to the user after filling in all the fields in the Login form in order to validate the registration.



## Welcome to GENECOMMITTEE

Once logged in, the user has access to the home page of GENECOMMITTEE, and is able to start a new work or continue a previously started project.

The email address of the active user is shown at the top of the page. Quick-navigation buttons can be found there, linking to the Help, user Personal Data, Committee Training, Diagnostic Mode, Data Management, Home and Logout pages.




guest\_1353266667158



The user must change his/her Email Notifications setting in the Personal Data tab in order to receive an email message after task completion. Note that email notifications will be sent to the active user, so you should create and use your own account to receive them.


GENECOMMITTEE is divided in three sub-tools: (i) Committee Training, (ii) Diagnostic Mode, and (iii) Data Management, as shown below.






### Committee Training

A six steps wizard will guide you in the creation and training of a new committee. Select a training data set, a set of gene sets and your preferred classifiers and form a new committee with the best experts (classifier & gene set)



### Diagnostic Mode

Use your committees to evaluate new patients. Each member of your committee will classify your patients using the knowledge adquired on its training. A final summary will be generated with useful diagnostic information



### Data Management

Upload and manage your training data sets

Prior to training any committees the user is required to upload the desired data sets using the Data Management screen. All uploaded data sets are stored in GENECOMMITTEE, allowing the user to suspend a project and continue it later without risk of losing previous work. Note that the demo account provides two default data sets for tool testing purposes. Once the data sets are uploaded, the user can directly proceed to the Committee Training tool and perform the desired experiments. All tasks presented in this guide were performed using the demo account and the <Valk> default data set.

## Committee Training



**Committee Training**

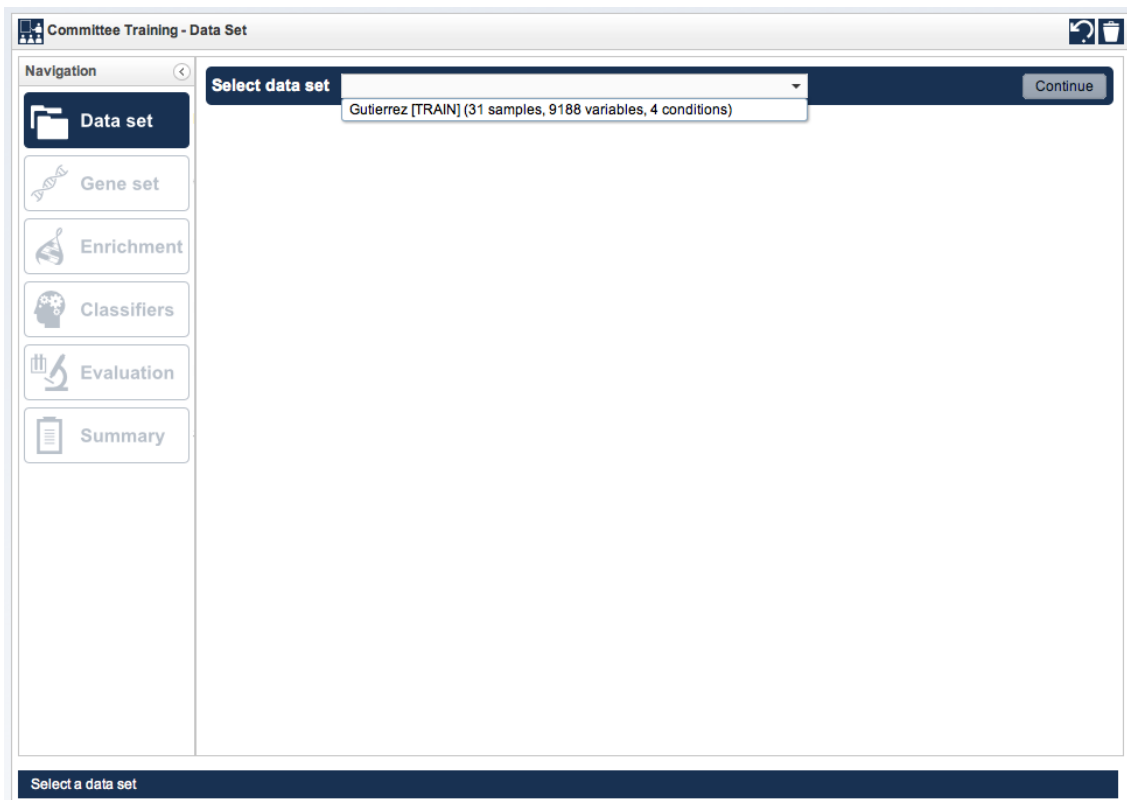
A six steps wizard will guide you in the creation and training of a new committee. Select a training data set, a set of gene sets and your preferred classifiers and form a new committee with the best experts (classifier & gene set)

The Committee Training wizard follows a workflow whose main steps are presented on the left toolbar.



**Data set** *Data Set Selection*

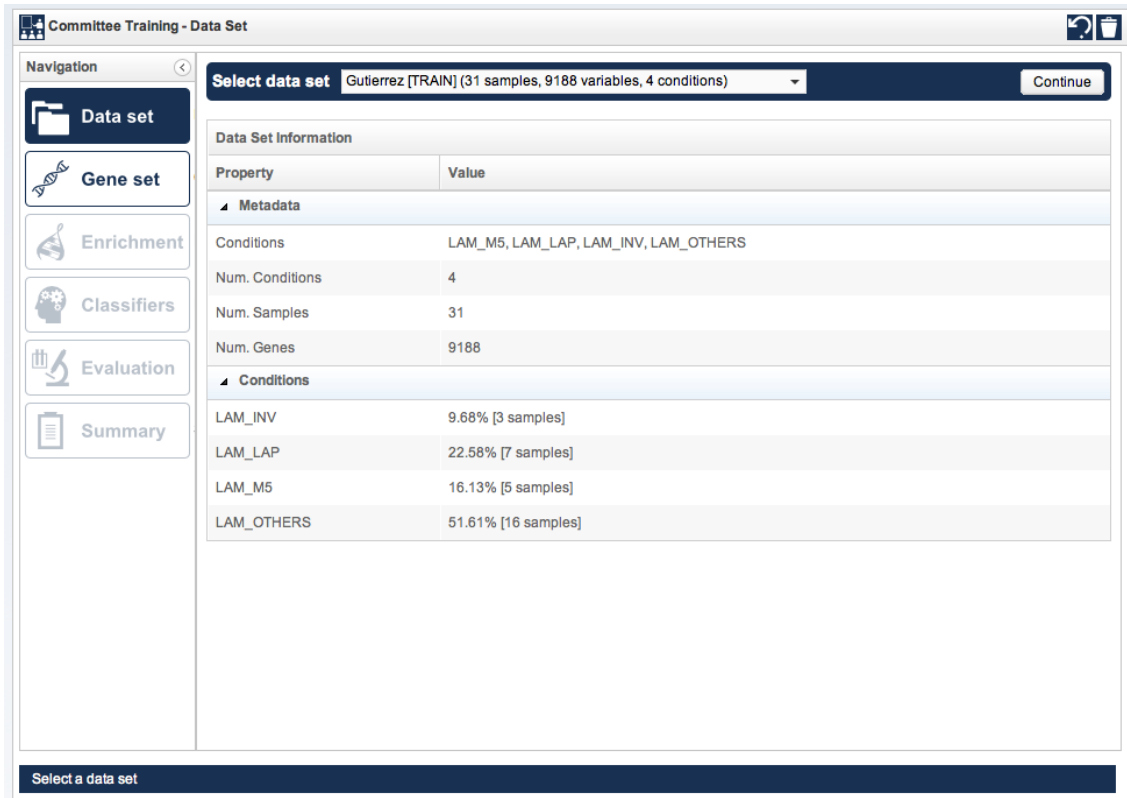
The first step consists of selecting the desired data set. To do so, click on the “Select data set” dropdown menu, choose the previously uploaded data set and click “Continue”.



The screenshot shows the 'Committee Training - Data Set' wizard interface. On the left is a navigation sidebar with buttons for 'Data set', 'Gene set', 'Enrichment', 'Classifiers', 'Evaluation', and 'Summary'. The 'Data set' button is highlighted. The main area features a 'Select data set' dropdown menu with a list of available datasets, including 'Gutierrez [TRAIN] (31 samples, 9188 variables, 4 conditions)'. A 'Continue' button is located to the right of the dropdown. At the bottom of the main area, there is a dark blue bar with the text 'Select a data set'.

Note that if the data set is changed during committee training, and before the whole workflow is completed and saved, the training process already performed upon the current dataset will be lost. Please make sure you select the correct data set, before proceeding.

For each data set, general information is shown relating to conditions and samples. Specific information properties can be expanded or collapsed. The information is presented in tabular view and is grouped in two tables: (i) Metadata (upper table), consisting of the sample size, number and discrimination of the conditions, and the number of genes present, and; (ii) Conditions (lower table), where each discriminated condition is represented.



The screenshot shows the 'Committee Training - Data Set' interface. The main area displays 'Data Set Information' for the selected data set 'Gutierrez [TRAIN] (31 samples, 9188 variables, 4 conditions)'. The information is presented in two tables: Metadata and Conditions.

Property	Value
<b>Metadata</b>	
Conditions	LAM_M5, LAM_LAP, LAM_INV, LAM_OTHERS
Num. Conditions	4
Num. Samples	31
Num. Genes	9188
<b>Conditions</b>	
LAM_INV	9.68% [3 samples]
LAM_LAP	22.58% [7 samples]
LAM_M5	16.13% [5 samples]
LAM_OTHERS	51.61% [16 samples]

The selected data set can then undergo gene selection.



The screenshot shows the 'Gene set' selection interface. It features a 'Gene set' button on the left and a 'Gene Selection' section on the right.

Using the dropdown menus in the blue bar shown below it is possible to filter the number of genes with higher discriminative ranking. The available options for gene selection are chi-squared distribution, information gain split method, gain ratio, and the relief-f feature filtering algorithm. Additionally, the user can opt to binarize numeric attributes and/or merge missing values in order to better adapt raw data to the selected filtering algorithm.

Once all the details are configured, it is necessary to press the “Select Genes” button to automatically perform the gene selection process.

The screenshot shows the 'Committee Training - Gene Set' interface. At the top, it says 'Select top 50 genes using Chi-square'. Below this is an 'Advanced Configuration' section with options for 'Binarize Numeric Attributes' and 'Missing Merge'. The main part of the interface is a table titled 'Selected Genes' with columns for Position, Gene, Ranking, and Info. The table lists 13 genes, with VCAN at the top (Ranking: 63.291666666666664) and HEXB at the bottom (Ranking: 49.26450216450217). A navigation bar at the bottom indicates '[FINISHED] Gene selection'.

Position	Gene	Ranking	Info
1	VCAN	63.291666666666664	
2	NRGN	61.999999999999999	
3	FLOT1	58.636284722222222	
4	ASNS	57.37605042016807	
5	CD14	54.766666666666666	
6	ITM2C	52.941558441558435	
7	CCND2	51.034013605442176	
8	FBP1	50.805555555555556	
9	SC5DL	50.73086734693879	
10	LILRB2	50.33226102941176	
11	RIN2	49.932142857142864	
12	FAM129A	49.332319078947364	
13	HEXB	49.26450216450217	

In the above example, five genes were selected using the information gain split method. Genes are ordered by test ranking.

In the following step of the workflow, the selected genes can be enriched using the integrated GeneBrowser tool.

**Enrichment**

***Enriched Gene Set***

The selected genes are transported to the Enrichment screen. Clicking the “Enrich” button results in a query to GeneBrowser and returns information regarding each gene. As soon as gene enrichment is completed, a new table containing the retrieved data is immediately shown.



Committee Training - Enrichment

Enrich your gene set using GeneBrowser Enrich Continue

Gene Sets Selected Gene Sets 27

Filters Name  Source  P-Value 0.0001 Coverage 25

Select All Unselect All Invert Selection Clear Filters

Selected	Name	Source	P-value	Coverage	Genes
<input checked="" type="checkbox"/>	P:cellular response to mechanical stimulus	Gene Ontology	0.0000	49	60
<input checked="" type="checkbox"/>	P:glycosphingolipid metabolic process	Gene Ontology	0.0000	25	38
<input checked="" type="checkbox"/>	P:positive regulation of tumor necrosis factor production	Gene Ontology	0.0001	25	42
<input checked="" type="checkbox"/>	P:cellular response to drug	Gene Ontology	0.0000	33	43
<input checked="" type="checkbox"/>	P:cellular defense response	Gene Ontology	0.0000	35	61
<input checked="" type="checkbox"/>	P:protein processing	Gene Ontology	0.0001	28	42
<input checked="" type="checkbox"/>	P:positive regulation of phosphatidylinositol 3-kinase	Gene Ontology	0.0000	37	49
<input checked="" type="checkbox"/>	P:regulation of cell shape	Gene Ontology	0.0000	77	120
<input checked="" type="checkbox"/>	P:defense response	Gene Ontology	0.0000	47	69
<input checked="" type="checkbox"/>	P:neutrophil chemotaxis	Gene Ontology	0.0001	32	41
<input checked="" type="checkbox"/>	P:myelination	Gene Ontology	0.0000	27	40
<input checked="" type="checkbox"/>	P:excretion	Gene Ontology	0.0001	32	45

[ 1 - 12 / 27 ]

[FINISHED] Gene set enrichment

All entries are selected by default, but the user has the possibility of choosing the desired enriched gene sets. Unwanted entries can be deselected using the checkbox in the first column.

The previous table contains the name of the found items, their sources, the associated  $p$ -value, the number of genes involved and the respective link to GeneBrowser<sup>1</sup>. This link allows the user to obtain detailed information for entries of interest without leaving GENECOMMITTEE.

Gene Explorer

0 - Gene Report for CAT

1 - Gene Report for VEGFA

2 - Gene Report for FAIM

Summary

Gene Name FAIM

Fullname Fas apoptotic inhibitory molecule 1

Synonyms FAIM1,

Function Plays a role as an inducible effector molecule that mediates Fas resistance produced by surface Ig engagement in B c

Location Cytoplasm

Uniprot Status Swiss-Prot

Gene Ontology

Biological Process GO:0006915 apoptotic process  
GO:0043066 negative regulation of apoptotic process

Molecular Function No data found

Cellular Component GO:0005737 cytoplasm

Pathway

Homologies

Structure

Sequence

References

3 - Gene Report for BARD1

4 - Gene Report for KRT18

5 - Gene Report for PROC

6 - Gene Report for NUP62

7 - Gene Report for SORT1

8 - Gene Report for ASNS

9 - Gene Report for ANGPTL4

10 - Gene Report for CLN3

11 - Gene Report for DFFA

12 - Gene Report for ALB

13 - Gene Report for ERCC5

<sup>1</sup> <http://bioinformatics.ua.pt/gb2>

Following gene set enrichment, it is necessary to configure the classifiers and the evaluation strategy.


Classifiers

Evaluation Configuration

The evaluation system in GENECOMMITTEE is very flexible, allowing the use of a single classifier or multiple classifier combinations. All classifiers have their advantages and drawbacks, so it is important to assign the most appropriate classifier to the data set in use. Our idea of committee allows choosing the best combination of methods for a specific data set.

There are five classifiers available: (i) k-nearest neighbours, (ii) decision trees, (iii) support vector machine, (iv) naïve Bayes, and (v) random forest. After being added, each classifier can be refined to perfectly assess the input data. This feature can be accessed by clicking the “Edit” button at the right of the classifier, just next to the “Delete” button. Also, the names of the classifiers can be changed for easier identification.

The last task consists of setting the evaluation strategy to finally perform the experiment. Note that clicking the “Continue” button will not promptly initiate the job, but will take the user to the Evaluation screen.

Evaluation

Experiment Execution

Once all the settings are defined, clicking the “Evaluate” button initiates the task. An attractive feature of GENECOMMITTEE is live-visualization of the experiment execution. If any error is detected, the user can cancel the active task clicking the “Abort” button in the progress bar.

The screenshot displays the 'Committee Training - Evaluation' window. It features a navigation sidebar on the left with options: Data set, Gene set, Enrichment, Classifiers, Evaluation (selected), and Summary. The main area is titled 'Evaluate your experts (classifier & gene set) and select your committee members' and contains a table of 'Execution Results'.

The table shows performance metrics for five classifiers across several gene sets. The classifiers are Decision Tree (C4.5), Naïve Bayes Simple, Random Forest, Support Vector Mach, and k-Nearest Neighbour. The gene sets listed are Asparagine synthetas, Versican (chondroitin), <Disease 1>, CD14 antigen, Neurogranin, Flotillin 1, asparagine biosynthe, phagocytosis, apoptosis, cell adhesion, signal transduction, cell surface receptor I, multicellular organisr, nervous system devel, and cell recognition.

Gene Set	Decision Tree (C4.5)	Naïve Bayes Simple	Random Forest	Support Vector Mach	k-Nearest Neighbour
Asparagine synthetas	0.6278	0.4624	0.5044	0.4019	0.4019
Versican (chondroitin)	0.4131	0.2991	0.3976	0.2530	0.2530
<Disease 1>	0.4131	0.2991	0.3976	0.2530	0.2530
CD14 antigen	0.4454	0.2855	0.2815	0.3542	0.3500
Neurogranin	0.3350	0.4280	0.3871	0.3613	0.3219
Flotillin 1	0.4105	0.1268	0.3374	0.2365	0.2000
asparagine biosynthe	0.6278	0.4624	0.5044	0.4019	0.4019
phagocytosis	0.6609	0.7919	0.7266		0.7608
apoptosis					
cell adhesion					
signal transduction					
cell surface receptor I					
multicellular organisr					
nervous system devel					
cell recognition					

At the bottom of the window, a status bar indicates: [RUNNING] Experiment execution (24.38% completed - 3 tasks running) [Abort]

As soon as the job is finished the user must pick the desired experts (classifier and gene features) for building a new committee. In order to evaluate the performance of each expert, the statistical analysis of the execution results can be adjusted. Allowed statistics include Cohen’s Kappa, accuracy, precision, recall, specificity and F-measure. This allows better perception of the results' significance.

Committee Training - Evaluation

Evaluate your experts (classifier & gene set) and select your committee members

Execution Results

Gene Set	Decision Tree (C4.5)	Naive Bayes Simple	Random Forest	Support Vector Machi	k-Nearest Neighbours
P:positive regulation	<b>0.6687</b>	<b>0.6146</b>	0.5507	0.2439	0.4259
P:positive regulation	<b>0.6511</b>	0.4249	0.5350	0.0000	0.4910
P:positive regulation	<b>0.6347</b>	0.5356	0.3099	0.3221	0.4697
P:regulation of cell	<b>0.6173</b>	<b>0.6004</b>	<b>0.6748</b>	0.1656	<b>0.8080</b>
P:cellular response	0.5463	0.5139	0.4384	-0.0464	0.2976
P:cellular defense	0.5428	0.5433	0.4058	0.3945	0.5245
P:defense response	0.5141	0.4665	0.5079	0.1725	0.5016
P:electron transport	0.4895	0.1956	0.4683	0.2470	0.5536
P:cellular response	0.4553	0.3638	0.2024	0.1656	0.4086
P:protein	0.4338	0.3505	0.4211	0.3248	<b>0.6137</b>
P:neutrophil	0.4188	0.4843	0.2584	0.0000	0.4241
P:regulation of blood	0.4142	0.4050	0.5579	0.2439	0.4692
P:response to	0.4038	0.4624	0.3060	0.2439	0.4746
P:excretion	0.3941	-0.1262	0.4048	0.0000	0.1533
P:myelination	0.3866	0.3366	0.4428	0.0844	0.5493

Execute the evaluation of the classifiers using selected gene sets

Another interesting feature is the possibility of visualizing the results of a specific class, or in the case shown, a condition. Note that results of all classes are shown by default. Statistic types and class visualization can be changed using the dropdown menus above the results table.

Once the experts are finally selected, the user must then save the information about the newly created committee.

Summary

**Committee Summary**

Finally, the Committee Summary screen shows general information about the input data set, how gene selection was performed and committee details. The newly created committee can be instantly used in the Diagnostic Mode to evaluate new patients.

Committee Training - Summary
🗑️

Navigation

Data set

Gene set

Enrichment

Classifiers

Evaluation

Summary


**Name your committee and finish**
Finish

Committee Name

Property	Value
<b>▲ Data Set</b>	
Name	Gutierrez [TRAIN]
Conditions	LAM_M5, LAM_LAP, LAM_INV, LAM_OTHERS
Num. Conditions	4
Num. Samples	31
Num. Genes	9188
<b>▲ Gene Selection</b>	
Name	Chi-square
Num. Genes	50
<b>▲ Committee</b>	
Num. Total Experts	135
Num. Selected Experts	10
Experts	Decision Tree (C4.5) # P:positive regulation of phosphatidylinositol 3-kinase cascade
	Decision Tree (C4.5) # P:positive regulation of protein phosphorylation
	Decision Tree (C4.5) # P:positive regulation of tumor necrosis factor production

[Check your generated committee](#)

## Diagnostic Mode



# Diagnostic Mode

Use your committees to evaluate new patients. Each member of your committee will classify your patients using the knowledge acquired on its training. A final summary will be generated with useful diagnostic information

In Diagnostic Mode the user can apply the created and trained committees to evaluate new patients. The carefully selected experts are compared to the test (patient) data to identify probable new disease cases.

Firstly, it is necessary to select the newly created committee using the dropdown menu in the blue bar. Then, patient data must be uploaded using the button on the right of the same bar.



The first time a user uses a committee for diagnostic, he/she should upload the patient data to be evaluated. Using the “Upload new patient data” button, the user can upload a data file with the patient data. This file follows the same format as described in the “Data Set Format” subsection but in this case, the “CLASS” row should not contain any value (it is recommended to use the “?” symbol as value).

Example patient data file:

UNIQUID	NAME	GSM20962.CEL	GSM20701.CEL	GSM20702.CEL
#	CLASS	?	?	?
1	HDAC1	0.26299268	0.45148313	-10.665.518
2	MAPRE1	0.16962652	-0.6115107	-0.5892962
3	NEO1	-0.80309725	-11.346.912	-12.663.207

As soon as the user uploads the patient data, the committee will start working on the diagnostic of the patients. The status bar at the bottom of the window shows you the diagnostic progress.

An information window will notify that the work has finished. When the diagnostic is complete, the user can explore the results by selecting the diagnostic in the second combo box of the upper tool bar. The diagnostic loading may take a while so please be patient while GENECOMMITTEE loads it.

**Diagnostic Mode**

Select a committee: Gutierrez [TRAIN] and view diagnostic: Gutierrez\_TEST.csv or Upload new patient data

**Gutierrez [TRAIN]**

Committee Info Rename Committee Delete Committee Rename Diagnostic Delete Diagnostic Incompatibilities Download Diagnostic Download Patient Data

Gutierrez_TEST.csv	AP13058_LAP	AP14398_LAP	AP5204_LAP	BP7644_INV
<b>Committee</b>				
Decision Tree (C4.5) # P:positive regulation of phosphatidylinositol 3-kinase cascade	LAM_LAP	LAM_LAP	LAM_LAP	LAM_OTHERS
Decision Tree (C4.5) # P:positive regulation of protein phosphorylation	LAM_LAP	LAM_LAP	LAM_INV	LAM_INV
Decision Tree (C4.5) # P:positive regulation of tumor necrosis factor production	LAM_LAP	LAM_OTHERS	LAM_LAP	LAM_INV
Decision Tree (C4.5) # P:regulation of cell shape	LAM_LAP	LAM_LAP	LAM_LAP	LAM_INV
Naïve Bayes Simple # P:positive regulation of phosphatidylinositol 3-kinase cascade	LAM_LAP	LAM_LAP	LAM_LAP	LAM_OTHERS
Naïve Bayes Simple # P:regulation of cell shape	LAM_LAP	LAM_LAP	LAM_LAP	LAM_OTHERS
Random Forest # P:regulation of cell shape	LAM_LAP	LAM_LAP	LAM_LAP	LAM_INV
k-Nearest Neighbours (1Bk) # P:phagocytosis	LAM_LAP	LAM_LAP	LAM_LAP	LAM_INV
k-Nearest Neighbours (1Bk) # P:protein homotetramerization	LAM_OTHERS	LAM_LAP	LAM_LAP	LAM_INV
k-Nearest Neighbours (1Bk) # P:regulation of cell shape	LAM_LAP	LAM_LAP	LAM_LAP	LAM_INV
<b>By Gene Set</b>				
P:phagocytosis	LAM_LAP	LAM_LAP	LAM_LAP	LAM_INV
P:positive regulation of phosphatidylinositol 3-kinase cascade	LAM_LAP	LAM_LAP	LAM_LAP	LAM_OTHERS
P:positive regulation of protein phosphorylation	LAM_LAP	LAM_LAP	LAM_INV	LAM_INV
P:positive regulation of tumor necrosis factor production	LAM_LAP	LAM_OTHERS	LAM_LAP	LAM_INV

[FINISHED] Evaluating patient

Diagnostic results are presented in a table where each column (except the first) represents one patient. In this table, the rows are grouped in four main sections:

- **Committee:** Each row contains the diagnostics of one member of the committee, a classifier trained using only the gene information of its associated gene set. Committee members will select one single condition for each patient.
- **By Gene Set:** This section summarizes the committee member's diagnostics by grouping the outputs of those members that share the same gene set. Only the condition or conditions with the highest number of votes are shown.
- **By Classifier:** In the same way as the previous section, this section groups the committee member's diagnostics by the classifier type employed.
- **Voting:** This final section summarizes the whole diagnostic process by showing the votes that each condition has received, along with a final row that shows the condition or conditions with the highest number of votes among all the committee members.

The diagnostic view also provides a helpful toolbar with several options that will help you to manage your trained committees and diagnostics. The options included in this toolbar are the following:

- **Committee Info:** Shows a popup window with information about the current committee. The figure below shows an example of this information panel.
- **Rename Committee:** This option allows you to rename the current committee.
- **Delete Committee:** This option allows you to delete the current committee. You must take into account that deleting a committee will provoke the deletion of the associated diagnostics.

- **Rename Diagnostic:** This option allows you to rename the current diagnostic. Diagnostics will be named by default with the name of the uploaded patient data set file.
- **Delete Diagnostic:** This option allows you to delete the current diagnostic.
- **Download Diagnostic:** With this option you can download the diagnostic information as a CSV file.
- **Download Patient Data:** This option allows you to download the original patient data set file.



## Data Management



In Data Management the user can upload and manage the data sets that will be used for committee training. As can be seen in the following figure, the Data Management view is divided into two main panels.

Variable	AP10222_S	AP12366_S	AP13223_S	AP14217_S	AP16089_S	AP16739_S	AP17074_S	BP185_S	BP355_S	BP10891_S	CP6667
Class	LAM_LAP	LAM_LAP	LAM_LAP	LAM_LAP	LAM_LAP	LAM_LAP	LAM_LAP	LAM_INV	LAM_INV	LAM_INV	LAM_M
SCN3A	-0.3899	-0.4425	-0.4572	-0.6335	-0.2451	0.2038	-0.1595	0.0545	0.5795	-0.2606	-0.416
SCN3B	-0.8545	-0.9782	-0.5732	-0.8949	-0.3751	0.3909	3.7660	-0.8241	-0.1870	0.1054	-1.607
GFER	0.1873	-0.5935	-1.2393	-1.6158	0.5212	-1.6345	0.1959	0.8990	-1.1953	-0.0802	0.115
RBM35A	-1.1663	-0.7928	-0.8238	-0.3409	-1.6072	0.7244	-0.1636	-0.2720	1.7629	-0.2271	-0.591
RBM35B	-0.2292	-0.9907	-2.5403	-0.2694	-1.7136	-1.3895	-0.1484	-0.3027	0.1693	-0.1501	0.928

The upper panel lists the existing data sets and allows the user to manage them. Using the tool included in the top toolbar, the user can search through the existing data sets or upload a new data set. Below this toolbar, a list of data sets shows the main features of each and allows the user to visualize, delete or download each data set.

When the user chooses to visualize a data set, it will be shown in the bottom panel. The data set will be displayed in a table view where the gene ids are listed in the first column, while the following columns contain the sample data. The first row contains the sample ids, the second row contains the class (condition) of each sample and the following rows contain the expression level of corresponding gene. In order to facilitate data set visualization, the upper panel can be collapsed (as shown in the following figure) using the top right button.

Data Set Management

Data Set Viewer

Variable	AP10222_S	AP12366_S	AP13223_S	AP14217_S	AP16089_S	AP16739_S	AP17074_S	BP185_S	BP355_S	BP10891_S	CP6667
Class	LAM_LAP	LAM_LAP	LAM_LAP	LAM_LAP	LAM_LAP	LAM_LAP	LAM_LAP	LAM_INV	LAM_INV	LAM_INV	LAM_M
SCN3A	-0.3899	-0.4425	-0.4572	-0.6335	-0.2451	0.2038	-0.1595	0.0545	0.5795	-0.2606	-0.416
SCN3B	-0.8545	-0.9782	-0.5732	-0.8949	-0.3751	0.3909	3.7660	-0.8241	-0.1870	0.1054	-1.607
GFER	0.1873	-0.5935	-1.2393	-1.6158	0.5212	-1.6345	0.1959	0.8990	-1.1953	-0.0802	0.115
RBM35A	-1.1663	-0.7928	-0.8238	-0.3409	-1.6072	0.7244	-0.1636	-0.2720	1.7629	-0.2271	-0.591
RBM35B	-0.2292	-0.9907	-2.5403	-0.2694	-1.7136	-1.3895	-0.1484	-0.3027	0.1693	-0.1501	0.928
BTBD2	0.3680	1.3210	-2.3592	-1.4179	-0.1587	0.6071	-0.6763	-0.6613	-1.6768	-0.1774	-0.335
BTBD3	0.3945	0.1710	2.1596	-0.2661	-0.7802	0.3738	-1.0571	0.4066	-0.7210	-0.8048	-0.591
TMEFF1	0.3992	-0.1183	-0.0276	2.1496	1.7748	0.9698	3.2979	-0.6590	0.4563	0.4960	0.484
SCN2B	0.0800	0.1034	-0.5971	0.7777	-3.8011	-0.6968	-2.2137	0.5009	-1.3101	-0.0285	1.215
TFPT	0.6330	0.8305	2.0030	-2.3307	-0.2980	0.8649	0.0314	0.2284	-0.9944	-0.8707	0.679
UGCG	-0.1552	-0.5203	0.0904	-0.3311	0.5728	1.9090	0.5562	-0.1486	1.3712	1.1328	-1.662
SCN2A	-0.4218	0.0235	0.8858	-0.3695	0.8502	-0.8564	0.9347	-1.5632	0.5634	0.7668	-1.823
BTBD7	0.5972	-1.6578	0.0216	1.4177	1.5038	-1.7720	2.7989	-0.8264	0.5387	-0.2050	-0.180
GEM	-0.4096	0.1317	-0.6622	2.2118	-1.7215	-1.1549	-1.0552	-0.8378	1.3288	-0.6672	-0.622
TAGLN	1.5769	1.5297	-0.9798	-0.3091	0.3752	-0.0822	1.1155	-0.1837	0.7000	-0.4827	0.506

[ 1 - 15 / 9187 ]



## Data Set Format

GENECOMMITTEE accepts comma separated value (CSV) files with the following structure:

- Samples are represented in the columns
- Genes are represented in the rows

Therefore, each sample (column) contains a:

- Sample ID;
- Class.

Each gene (row) contains a:

- Numerical identifier; Note that this gene identifier is not a database identifier, only a unique value to identify the gene in the data set;
- Name.

Each cell in the samples x genes matrix consists of the expression value of each gene.

These characteristics will require a specific input file format. The first row contains:

- First column: "UNQID";
- Second column: "NAME";
- Other columns: sample identifiers.

The second row contains:

- First column: "#";

- Second column: “CLASS”;
- Other columns: class names.

Other rows contain:

- First column: a numerical value different in all lines to identify each gene;
- Second column: gene names;
- Other columns: gene expression values (decimal, separated by a dot) for each sample, one column per sample.

Example file:

UNIQID	NAME	GSM20962.CEL	GSM20701.CEL	GSM20702.CEL
#	CLASS	LAM_RESTO	LAM_INV	LAM_INV
1	HDAC1	0.26299268	0.45148313	-10.665.518
2	MAPRE1	0.16962652	-0.6115107	-0.5892962
3	NEO1	-0.80309725	-11.346.912	-12.663.207