SensSpec.exe:

An algorithm created to cluster spoligotypes and MIRU-VNTRs and calculate sensitivity and specificity

1. Utility of the program

This algorithm was developed to estimate the sensitivity and specificity of spoligotyping and MIRU-VNTR typing using IS6110 RFLP typing as the reference standard (as referenced in Scott *et al* 2005). To do this, the algorithm compares the MIRU-VNTRs/spoligotypes of all the isolates in a dataset, determines the number of differences between each MIRU-VNTR/spoligotype, and creates the MIRU-VNTR/spoligotype clusters. The output lists each isolate, the number of the cluster it was assigned to, and the isolates it was clustered with. Descriptive statistics are provided (ie. the number of clustered isolates, the mean number of isolates per cluster, the size of each cluster etc), as well as the sensitivity/specificity. A special feature of this algorithm is that it will create not just clusters based on identical patterns, but also clusters allowing any number of differences between patterns (analogous to IS6110 RFLP clusters allowing for a band addition, deletion or shift). Therefore, this algorithm can be used not only to estimate sensitivity and specificity, but also to perform rapid clustering analysis of spoligotypes and MIRU-VNTRs.

2. How clusters are computed

MIRU-VNTR

The algorithm compares the first MIRU-VNTR in the database to every other MIRU-VNTR in the database, noting and recording how many loci are different for each comparison. All MIRU-VNTRs that match the first isolate (according to pre-set criteria) are considered to be a cluster and are assigned a cluster number. The algorithm then compares the MIRU-VNTR pattern of the second isolate in the database to the rest of the patterns, and any isolates with matching patterns are assigned a cluster number, and so on. During the clustering process the algorithm checks the assigned clusters to determine if any isolates in one cluster match isolates in another cluster. If matches between clusters are identified, the clusters are consolidated.

It is possible to vary the criteria for clustering to allow as many differences as desired between isolates. For instance, instead of clustering just isolates with identical copy numbers in each MIRU-VNTR locus, one could calculate clusters allowing one locus to differ (ie. a MIRU-VNTR pattern within a cluster would have the same number of copies at 11/12 loci to at least one other pattern). This information is entered in the user interface (see part 6).

Spoligotyping

The procedure for computing spoligotype clusters is the same as for MIRU-VNTR. For two spoligotypes to be identical, the algorithm requires every unique spacer present in one isolate to also be present in the other.

The criteria for clustering can be varied for spoligotype clusters, similar to what is available for MIRU-VNTR. However, variation in the DR locus can occur by two different methods: recombination between direct repeats or IS6110 sequences, and disruption of a single direct variable repeat (DVR) by insertion of IS6110 (Warren *et al* 2002). Therefore, this algorithm has been designed to calculate clusters two different ways.

1) A deletion of contiguous DVRs is assumed to be due to a single recombination event between direct repeats, and is therefore counted as one difference. In the program interface this is referred to as "Contiguous missing spacers = one deletion".
2) Each deletion/disruption of a DVR is considered independently of any other DVR, therefore each missing spacer is considered to be one difference. This is referred to as "Each missing spacer = one deletion" in the program interface.

"Each missing spacer = one deletion".

When calculating clusters based on independent spacers, the algorithm counts every spacer that is present in one isolate but not in the other and sums up the total number of spacers that are different. If this number is less than or equal to that allowable by the chosen clustering criteria, the isolates are considered matched and are assigned a cluster number. For example, isolates 1 and 2 would have 2 differences, isolates 2 and 3 would have 4 differences.

isolate 1: ■■■■■■■■■■■■■■■■■■■■••••■■■■■■■■••••■■■■■■■
isolate 2: ■■■■■■■■■■■■■■■■■■■■••■■■■■■■■■■••••■■■■■■■
isolate 3: ■■■■■■■■••■■■■■■■■■■■■■■■■■■■■■■■••••■■■■■■■
(Note: ■ represents spacers that are present, • represents absent spacers)

"Contiguous missing spacers = one deletion".

When calculating contiguous deletions as genetic events, the algorithm scans the spoligotypes, noting the position and the number of blank spacers in a row. A set of contiguous, blank spacers are considered one genetic event. When calculating the number of genetic event differences between two isolates, the program takes into account that an additional deletion could have occurred to extend the original deletion by any number of spacers on one side, the other, or both.

isolate 4: ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■••••■■■■■■■
isolate 5: ■■■■■■■■■■■■■■■■■■■■••••■■■■■■■■••••■■■■■■■
isolate 6: ■■■■■■■■■■■•••••••••••••••••■■■■••••■■■■■■■
isolate 7: ••••••••••••••••••••••••••••••••••■■■■■■■■■□

For instance, the program would count one genetic event difference between isolate 4 and isolate 5. However, there is also one genetic event difference between isolates 5 and 6, because the direct repeat 5' (to the left) of spacer 12 could have recombined with the direct repeat 3' (to the right) of spacer 38, deleting out all the intervening spacers. Sample 4 and 7 would have at least two genetic events because the common ancestor would have to look like:

■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■••■■■■■■■■■

The deletion of spacers 35 and 36 would lead to spoligotype 4, while deletion of spacers 1 – 34 would lead to spoligotype 7.

The number of differences between isolates are noted, and isolates with less than or equal to the number of differences set as the clustering criteria are assigned a cluster number and considered clustered.

3. How sensitivity and specificity are computed

Specificity

Specificity is defined as the probability of testing negative, given that the patient does not have the disease. In tuberculosis molecular epidemiology terms, this would be the probability of an isolate being called unique by MIRU-VNTR, given that it is considered unique by IS6110 RFLP. The algorithm uses the following formula:

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives + number of false positives}}$$

$$= \frac{\text{number of isolates not clustered by MIRU-VNTR}}{\text{all isolates not clustered by IS}\mathit{6110}}$$

Therefore, to obtain a measure of specificity, the database entered into the program must only contain isolates considered unique by IS6110 RFLP.

Sensitivity

Sensitivity is defined as the probability of testing positive given that the patient has the disease. For our purposes we can restate this to read: the probability of an isolate clustering by MIRU-VNTR given that the isolate is clustered by IS*6110* RFLP. The formula is therefore:

$$\text{Sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives + number of false negatives}}$$

$$= \frac{\text{isolates clustered by MIRU-VNTR within IS6110 clusters}}{\text{all isolates clustered by IS}\mathit{6110}}$$

Sensitivity is estimated in the population of isolates clustered by IS*6110* RFLP. However, an isolate is only compared to other isolates within the same IS*6110* cluster. If the algorithm compared isolates between IS*6110* clusters, then there would be comparisons (and possible matches) between isolates that are not supposed to be clustered. This could over-inflate the sensitivity estimates. For instance, isolates 24, 100 and 45 would be compared to each other, but 24 and 95, or 24 and 34, or 95 and 68 would not.

IS*6110* cluster	Isolate ID#	Spoligotype
1	34	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■••••■■■■■■■
1	68	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■••••■■■■■■■
2	95	■■■■■■■■■■■■■■■■■■■■••••■■■■■■■■••••■■■••••
2	67	■■■■■■■■■■■■■■■■■■■■••••■■■■■■■■••••■■■••••
3	24	■■■••••••••••■■■■•■■■■■■■■■■■■■■■■••••■■■■■■■
3	100	■■■••••••••••■■■■•■■■■■■■■■■■■■■■■••••■■■■■■■
3	45	■■■••••••••••■■■■•■■■■■■■■■■■■■■■■••••■■■■■■■

To estimate sensitivity, the database entered into algorithm must only contain isolates clustered by IS6110 RFLP, and the IS6110 cluster number must be indicated.

4. Installation ("INSTALL")

Once you have unzipped SensSpec.zip, the program SensSpec.exe (the only one you will need to call directly)is ready to use.

You will notice that a subdirectory called "src/" was created when you unzipped SensSpec.zip; important files are saved in this directory and it is important that they are neither deleted nor changed. When running SensSpec.exe, temporary files will also be saved in a subdirectory called tmp, where info regarding last input files analysed and your favorite printout settings will be saved. That directory will be created whenever you run SensSpec.exe, if deleted.

The only package you need to have on your computer to run SensSpec.pl is Active Perl (freely distributed: http://www.activestate.org). Once it is installed, make sure that Perl is in your path. Assuming Perl was installed in c:\Perl\bin, you need to add c:\Perl\bin to your PATH variable (change this accordingly if you installed Perl in a different directory).

To add to your PATH proceed as follows:

* If using Windows 2000/NT: click right mouse button on "My computer" and select "properties" then "Advanded" and "Environment variables" and edit PATH variable there.
* If using Windows 95/98: edit autoexec.bat file and modify PATH variable there. With Windows 95, make sure that the folder names do not contain blanks, but instead use the short DOS name (for example c:\progra~1\... in stead of c:\program files\...).

Modifying PATH will make the command perl.exe [used internally in SensSpec.exe] available.

View output file online:

Shall you want to have a glance at the output files produced by this program by just one click, you will also need to add the path to winword.exe to your PATH variable. However, if you decide not to this, SensSpec.exe will still produce the output, and you will still be able to view the outputs produced by browsing Windows Explorer, for example.

5. Data file preparation

MIRU-VNTR

For use in this algorithm, it is convenient to store MIRU-VNTR data in a Microsoft Excel spreadsheet, using one column per locus. The column titles must appear in the first row, and unless estimating sensitivity, the Patient ID number must be in the first column, with the MIRU-VNTR data immediately after. For example:

PATIENT	MIRU2	MIRU4	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40
25	2	2	6	4	2	5	1	5	3	3	2	2
58	2	2	5	3	3	5	1	5	3	3	1	3
157	2	2	5	2	1	3	2	4	6	2	2	1
496	2	2	5	3	2	4	1	5	3	3	3	1
697	2	2	4	3	1	3	1	5	2	3	2	1

However, the datafile must be converted to a tab deliminated text file before being entered into the program. An Excel spreadsheet can be easily converted into a tab deliminated text file, using the "Save as" function in Excel. The resulting file will appear thus:

```
PATIENT     MIRU2   MIRU4   MIRU10  MIRU16  MIRU20  MIRU23  MIRU24  MIRU26  MIRU27  MIRU31  MIRU39  MIRU40
25   2      2       6       4       2       5       1       5       3       3       2       2
58   2      2       5       3       3       5       1       5       3       3       1       3
157  2      2       5       2       1       3       2       4       6       2       2       1
496  2      2       5       3       2       4       1       5       3       3       3       1
697  2      2       4       3       1       3       1       5       3       3       2       1
```

Specificity. To estimate specificity of MIRU-VNTR using IS6110 RFLP as the reference (gold) standard, the datafile can only contain isolates that are considered unique by IS6110 RFLP (ie. are not part of an IS6110 cluster). The datafile would look identical to the ones demonstrated above.

Sensitivity. To estimate the sensitivity of MIRU, the datafile can only contain the isolates clustered by IS6110 RFLP, a cluster number must be given to each IS6110 cluster, and that IS6110 cluster number must be indicated in the datafile. **The first column must be the IS6110 cluster number, and start with the word "cluster" in the title.** The Excel spreadsheet would appear like so:

Cluster #	PATIENT	MIRU2	MIRU4	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40
1	853	2	2	6	3	2	5	1	5	3	3	2	3
1	4871	2	2	5	3	2	5	1	5	3	3	2	3
2	676	2	2	5	3	1	3	1	4	3	3	2	1
2	94322	2	2	5	3	1	4	1	5	3	3	2	1
2	3324	2	2	4	3	1	3	1	5	3	3	2	1

The tab deliminated text file would be:

Cluster #	PATIENT	MIRU2	MIRU4	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40
1	853	2	2	6	3	2	5	1	5	3	3	2	3
1	4871	2	2	5	3	2	5	1	5	3	3	2	3
2	676	2	2	5	3	1	3	1	4	3	3	2	1
2	94322	2	2	5	3	1	4	1	5	3	3	2	1
2	3324	2	2	4	3	1	3	1	5	3	3	2	1

Cluster Comparison Only. If the algorithm is being used to compute MIRU-VNTR clusters only, the datafile would contain all isolates during the period of interest, regardless of IS6110 RFLP results. The file would look identical to that used for specificity.

Spoligotyping

Spoligotyping data is also stored in a Microsoft Excel spreadsheet, but with the spoligotype within one column. Unless estimating sensitivity, the patient ID must be in the first column, with spoligotype data immediately after. Column titles must appear in the first row.

PATIENT	SPOLIGO
36789	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■•■••••■■■■■■■□
5867	■■■■■■■■■■■■•■■■■■■■■■■■■■■■■•■••••■■■■■■■□
357	■■■■■■■■■■■■■■■■■■■■■■■■■••••••■••••■■■■■■■□
32	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■•■••••■■■■■■■□
97543	■■■■■■■■■■■■■■■■■•■■■■■■■■■■■■•■••••■■■■■■■□

In the tab-deliminated text file, the symbol for spacers that are present (here □) must convert to a "g", and the symbol for absent spacers (in this document □) must convert to an "i". If your raw spoligotype data is entered using the Marlett font, this should automatically convert to the appropriate format. If not, the "Find and Replace" function of Microsoft Excel or Microsoft Word can easily convert your data into the required format.

```
PATIENT     SPOLIGO
36789    ggggggggggggggggggggggggggggggigiiiiggggggg
5867     ggggggggggggigggggggggggggggggigiiiiggggggg
357      gggggggggggggggggggggggggiiiiiigiiiiggggggg
32       ggggggggggggggggggggggggggggggigiiiiggggggg
97543    gggggggggggggggggiggggggggggggigiiiiggggggg
```

Specificity. As mentioned above, to estimate specificity of spoligotyping, the datafile can only contain isolates that are considered unique by IS6110 RFLP. The datafile would look identical to the one above.

Sensitivity. To estimate spoligotyping sensitivity, the datafile can only contain those isolates clustered by IS6110 RFLP and the IS6110 cluster number must be indicated in the first column. **The title of the first column must start with the word "cluster".** The Excel spreadsheet would appear thus:

Cluster #	PATIENT	SPOLIGO
1	4577	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■•■••••■■■■■■■□
1	785	■■■■■■■■■■■■•■■■■■■■■■■■■■■■■■■■■•■••••■■■■■■■□
2	327	■■■■■■■■■■■■■■■■■■■■■■■■■••••••■••••■■■■■■■□
2	5799	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■•■••••■■■■■■■□
2	4257	■■■■■■■■■■■■■■■■■•■■■■■■■■■■■■•■••••■■■■■■■□

The resulting tab-deliminated text file:

```
Cluster #   PATIENT     SPOLIGO
1      4577  ggggggggggggggggggggggggggggggigiiiiggggggg
1      785   ggggggggggggigggggggggggggggggigiiiiggggggg

2      327   ggggggggggggggggggggggggggiiiiiigiiiiggggggg
2      5799  ggggggggggggggggggggggggggggggigiiiiggggggg
2      4257  gggggggggggggggggigggggggggggggigiiiiggggggg
```

Cluster computation only. If one is not interested in sensitivity/specificity, and only wants to calculate the spoligotype clusters, the datafile would contain all isolates during the period of interest, regardless of IS6110 RFLP results. The file would look identical to that used for specificity.

6. How to use the interface (“README”)

The main program of this package is SensSpec.exe: it is the program that you will use (by clicking on it in Windows Explorer) to analyze subjects' strains, saved beforehand in a text file.

Before using SensSpec.exe, please read the file help\INSTALL.doc and follow the short instructions.

Once launched, SensSpec.exe will pop up a graphical user interface in which you will have to select the input file, define the criterion for clustering, specify the level for confidence interval to be reported and specify an output file name.

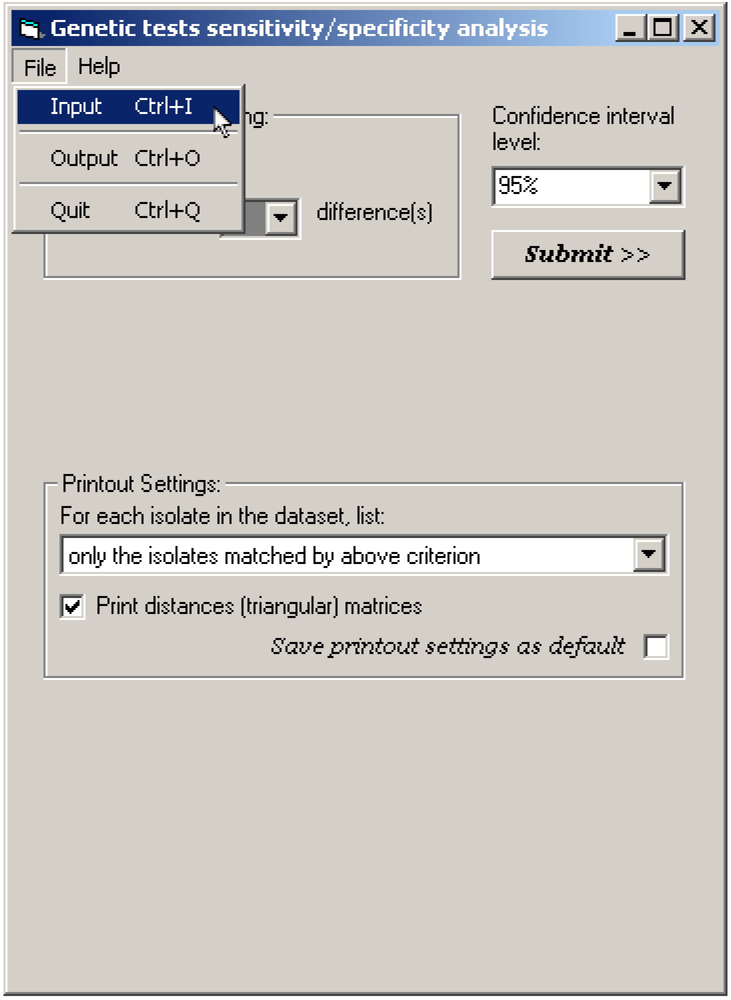
Guided tour to SensSpec.exe

We present here the steps to follow to perform an analysis with SensSpec.exe. Points are presented in a natural order, but feel free to do it in the order that comes to your mind, as the program will not complain.

Select input file

In the tool bar, select File File and Input (or Ctrl+I as a shortcut).

That will prompt an Open Dialog window (below), where you will be able to select your input file.



difference(s)

Confidence interval level: 95%

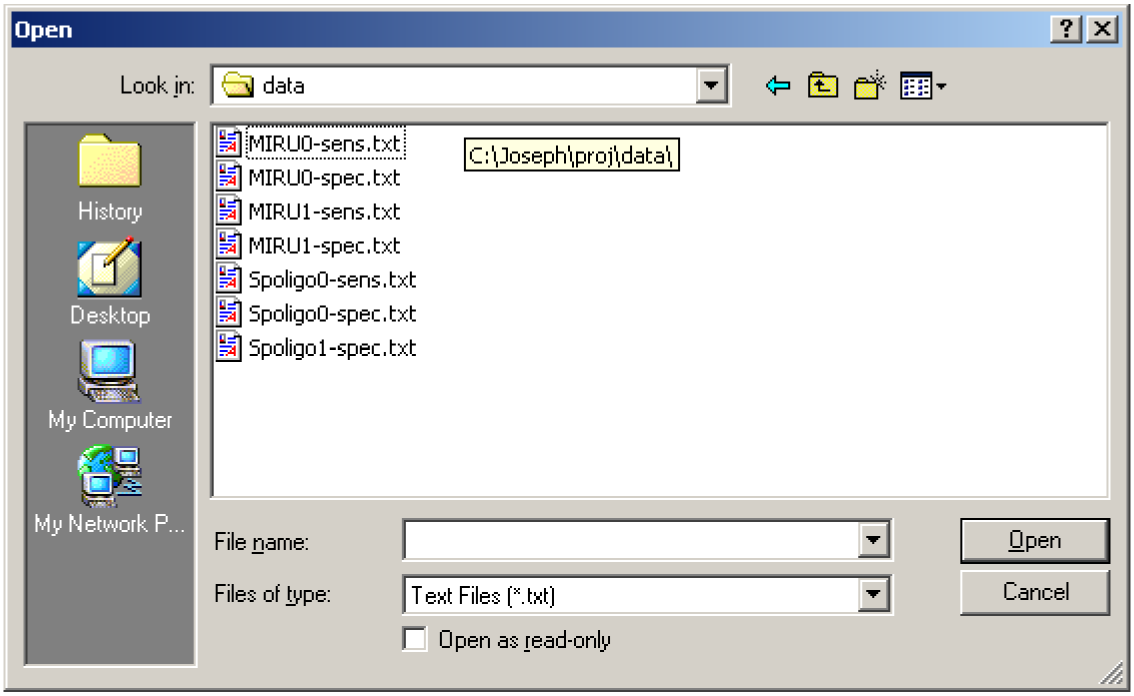
Printout Settings:

For each isolate in the dataset, list: only the isolates matched by above criterion

☑ Print distances (triangular) matrices

Save printout settings as default □

Browse through your folders to select your input file.



Look in: data

MIRU0-sens.txt
MIRU0-spec.txt
MIRU1-sens.txt
MIRU1-spec.txt
Spoligo0-sens.txt
Spoligo0-spec.txt
Spoligo1-spec.txt

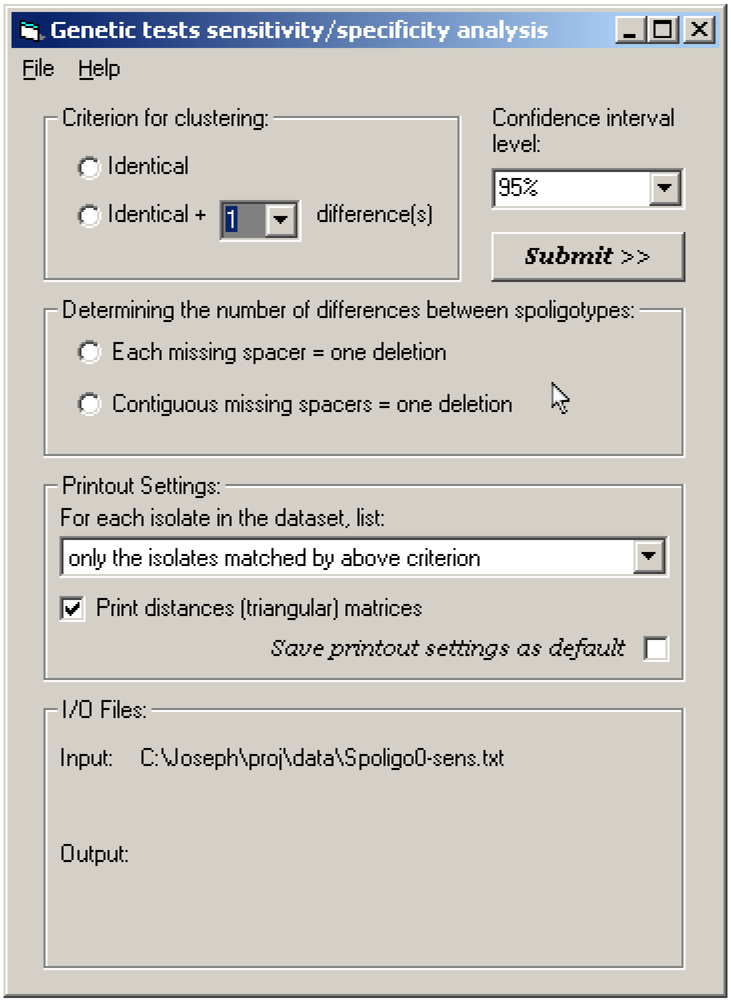
C:\Joseph\proj\data\

File name:

Files of type: Text Files (*.txt)

□ Open as read-only

Notice that a new frame appeared at the very bottom part of the form: it is a reminder of which file was selected as the input file and, eventually, of the location of the output file.



Criterion for clustering:

○ Identical

○ Identical + 1 difference(s)

Confidence interval level: 95%

Determining the number of differences between spoligotypes:

○ Each missing spacer = one deletion

○ Contiguous missing spacers = one deletion

Printout Settings:

For each isolate in the dataset, list: only the isolates matched by above criterion

☑ Print distances (triangular) matrices

Save printout settings as default □

I/O Files:

Input: C:\Joseph\proj\data\Spoligo0-sens.txt

Output:

Spoligotype

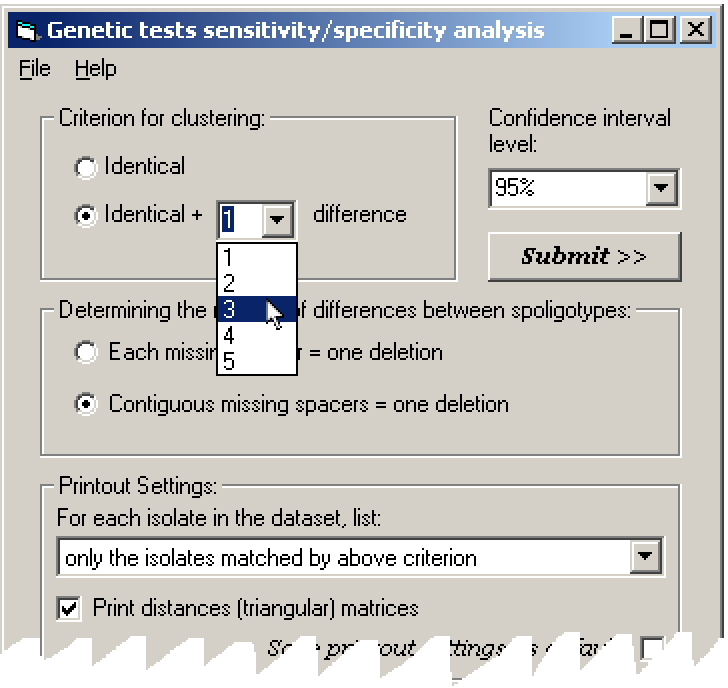
When input file contains spoligotype data, differences between two strains can be computed in two ways. Click the cell corresponding to your choice (second frame, starting from top).

Selecting confidence interval level

Confidence interval level can be selected through the list box in upper right corner of the form. Possible values are 80%, 90%, 95% and 99%; these choices are offered for the sake of completeness, but it is common practice to report 95% confidence intervals.

Criterion for clustering

When building clusters, you can regroup in a cluster only subjects that are identical to each other, or regroup patients that have at most some predetermined number of differences (*n*) with at least another subject of the cluster; if your criterion is the former, click Identical, while if it is the latter, click the cell at the left of Identical+ and specify the value for *n* in the cell at the right of Identical+: you can either pick one of the values offered in the list box at the right of Identical+ (from 1 to 5) or enter any other value yourself in the box.



Criterion for clustering:

○ Identical

● Identical + 1 difference

Confidence interval level: 95%

Determining the number of differences between spoligotypes:

○ Each missing spacer = one deletion

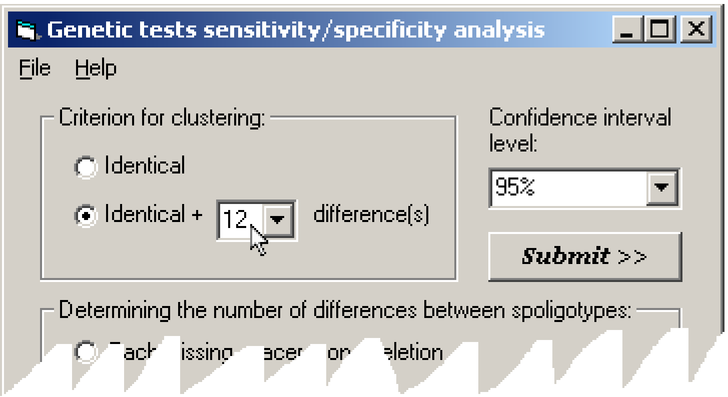
● Contiguous missing spacers = one deletion

Printout Settings:

For each isolate in the dataset, list: only the isolates matched by above criterion

☑ Print distances (triangular) matrices

Entering a value for *n* larger than 5. Click in the cell to the right of Identical+ and enter the value for *n*.


Genetic tests sensitivity/specificity analysis
File
Help
Criterion for clustering:
○ Identical
● Identical + 12 difference(s)
Confidence interval level:
95%
Submit >>
Determining the number of differences between spoligotypes:
○ Each issing acer on eletion

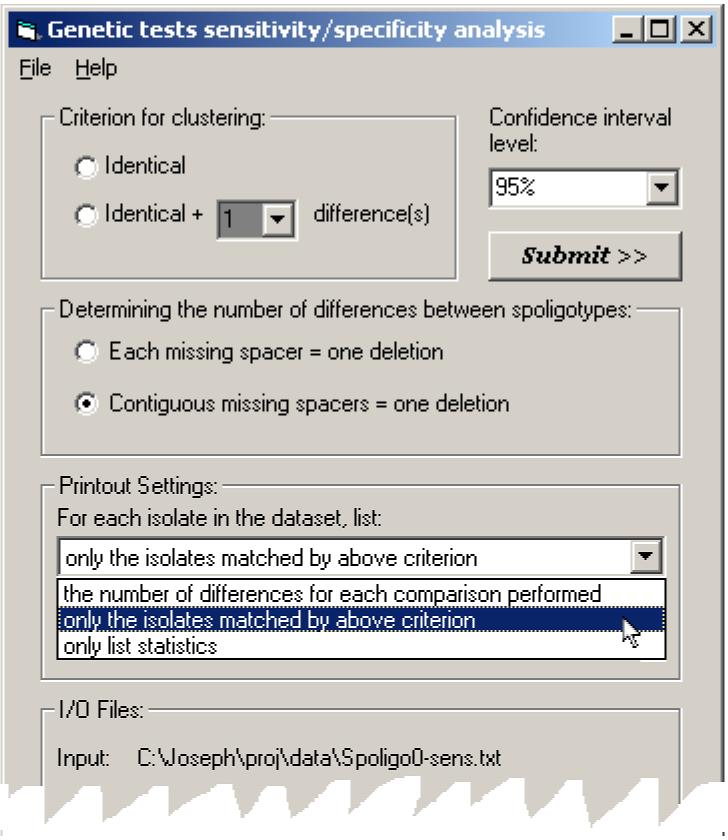
Printout settings

You have some control on what is going to be printed in the output file.

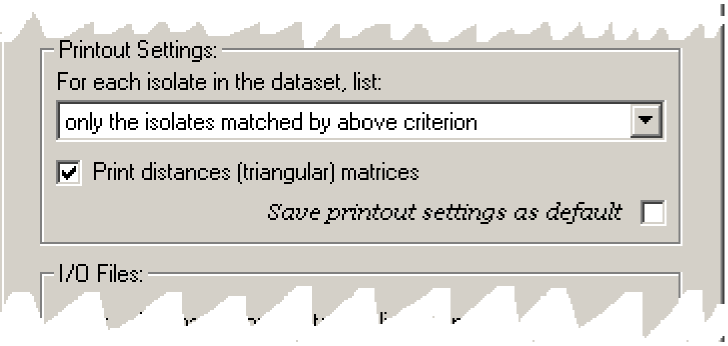
In the Printout Settings frame, you can list, for each subject, the complete sorted list (by increasing number of differences with the former) of subjects in the same original cluster (or in the complete data set if data do not originally include cluster numbers) by choosing the first option in the list box under "For each isolate in the data set, list:".

Conversely, one can list only the differences between subjects in the same new cluster (as defined by the criterion specified earlier) by selection the second option of the list box.

And, finally, one may prefer not to have the list of differences printed in the report; click "only list statistics".


Genetic tests sensitivity/specificity analysis
File
Help
Criterion for clustering:
○ Identical
○ Identical + 1 difference(s)
Confidence interval level:
95%
Submit >>
Determining the number of differences between spoligotypes:
○ Each missing spacer = one deletion
● Contiguous missing spacers = one deletion
Printout Settings:
For each isolate in the dataset, list:
only the isolates matched by above criterion
the number of differences for each comparison performed
only the isolates matched by above criterion
only list statistics
I/O Files:
Input: C:\Joseph\proj\data\Spoligo0-sens.txt

A generally more concise way to get the differences between subjects is to print a distances matrix. It simplifies your task when you are interested in the number of differences between the strains of two subjects in particular.

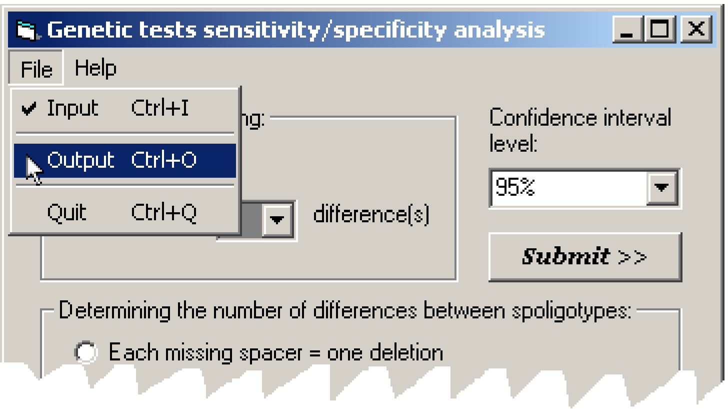

Printout Settings:
For each isolate in the dataset, list:
only the isolates matched by above criterion
☑ Print distances (triangular) matrices
Save printout settings as default □
I/O Files:

We have found useful, in practice, to print the distances matrix along with the reduced list of comparisons (by selecting "only the isolates matched by the above criterion" above).

Finally, you might want to save your printout settings options for future use by clicking the cell at the bottom right corner of the Printout Settings frame.

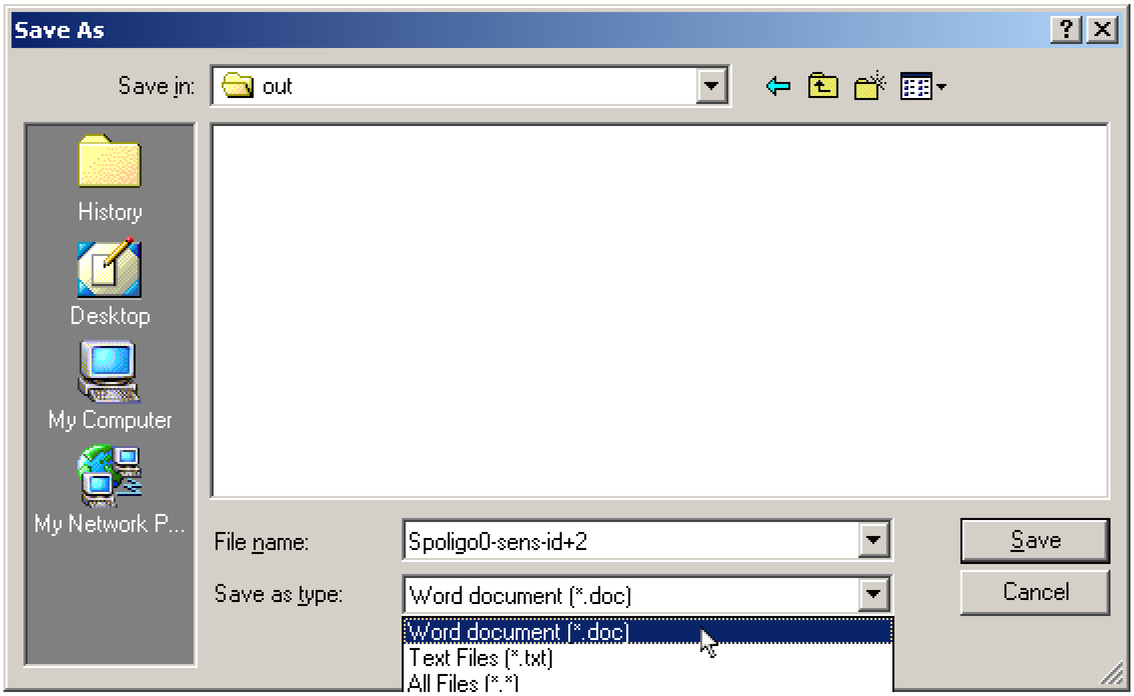
Selecting output destination

Prior steps have fully described the problem; the only thing left before submitting the problem is to select the output destination.


Genetic tests sensitivity/specificity analysis

Determining the number of differences between spoligotypes:
○ Each missing spacer = one deletion

In the tool bar, click File and Output, or type Ctrl+O as a shortcut.



Save in: out

File name: Spoligo0-sens-id+2

Save as type: Word document (*.doc)

Word document (*.doc)
Text Files (*.txt)
All Files (*.*)

An Open Dialog window (right) will open. Browse through your folders to select the output destination.

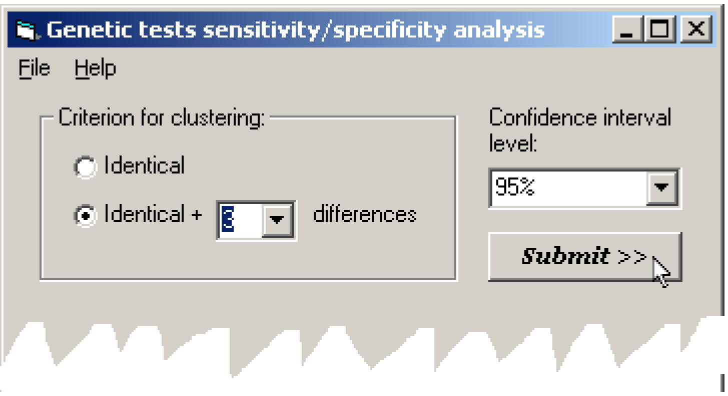
You can click the name of a pre-existing file* or enter a new file name (in the File name box).

*Note that any pre-existing file would be overwritten without asking for confirmation.

Note that the output can be saved in Word format (.doc), as plain text file (.txt) or with any extension of your choice.

Submitting the program

Submit the program by clicking the command button Submit, in the top right corner of the form.



Criterion for clustering:

○ Identical

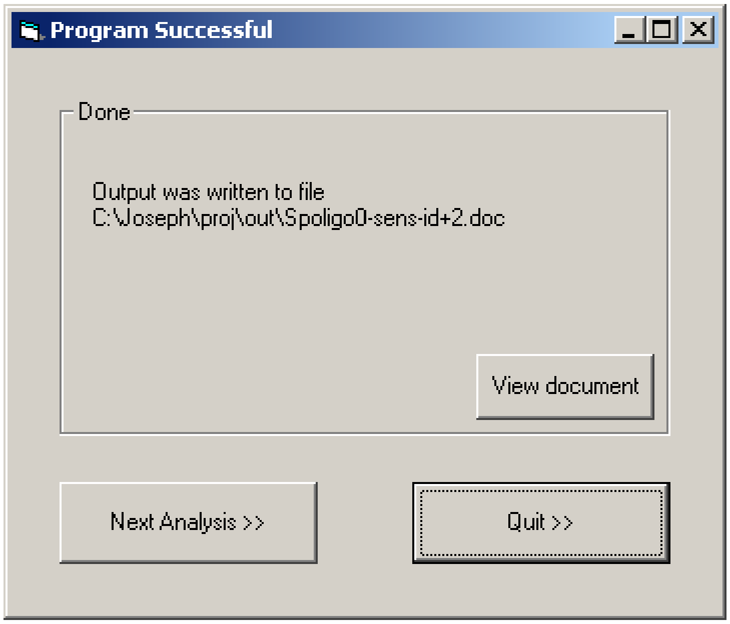
● Identical + 3 differences

Confidence interval level: 95%

Success

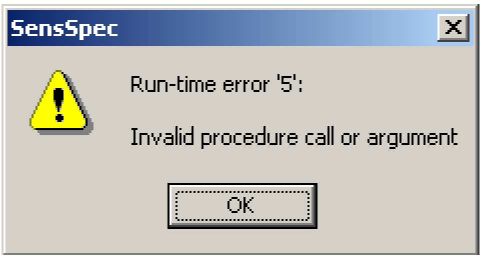
If the program is successful, a Program Successful window will be prompted. It reminds you where the output file was saved and offers you the possibility to view it immediately in Word (that options needs some attention at the installation of this package, though: see help\Install.doc for details).

Other options are to quit or to proceed to a new analysis.



Output was written to file
C:\Joseph\proj\out\Spoligo0-sens-id+2.doc

When you click View document to view the output file immediately, a Run-time error message will pop up if Word was already open. However, the document should still open successfully: thus, you can ignore this error message.



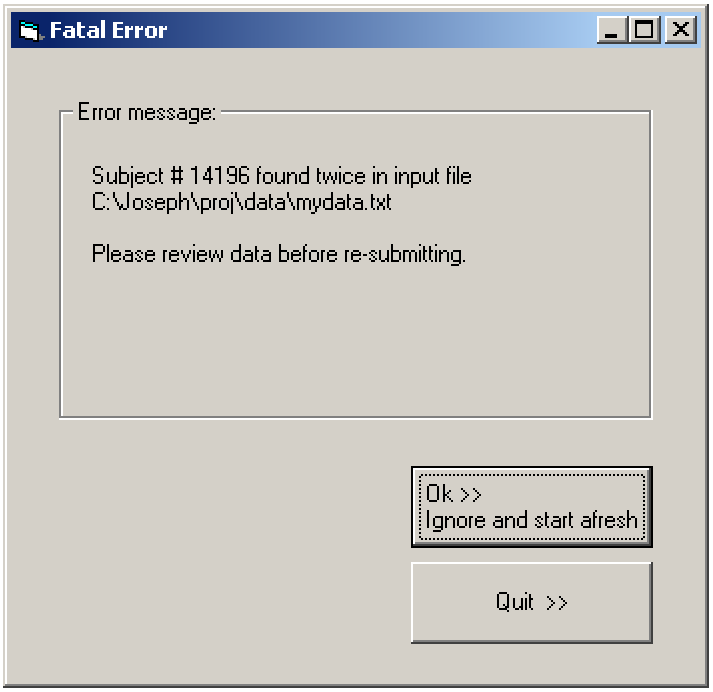
Run-time error '5':

Invalid procedure call or argument

Fatal error

Some mistakes in the data files can lead to fatal errors, making SensSpec.exe unable to perform clustering. Such fatal errors will be displayed in a proper window with a message that we hope to be helpful in tracking down the source of the problem.

Your options will be to quit or to ignore the error message and the analysis you were attempting to do and proceed with another analysis.

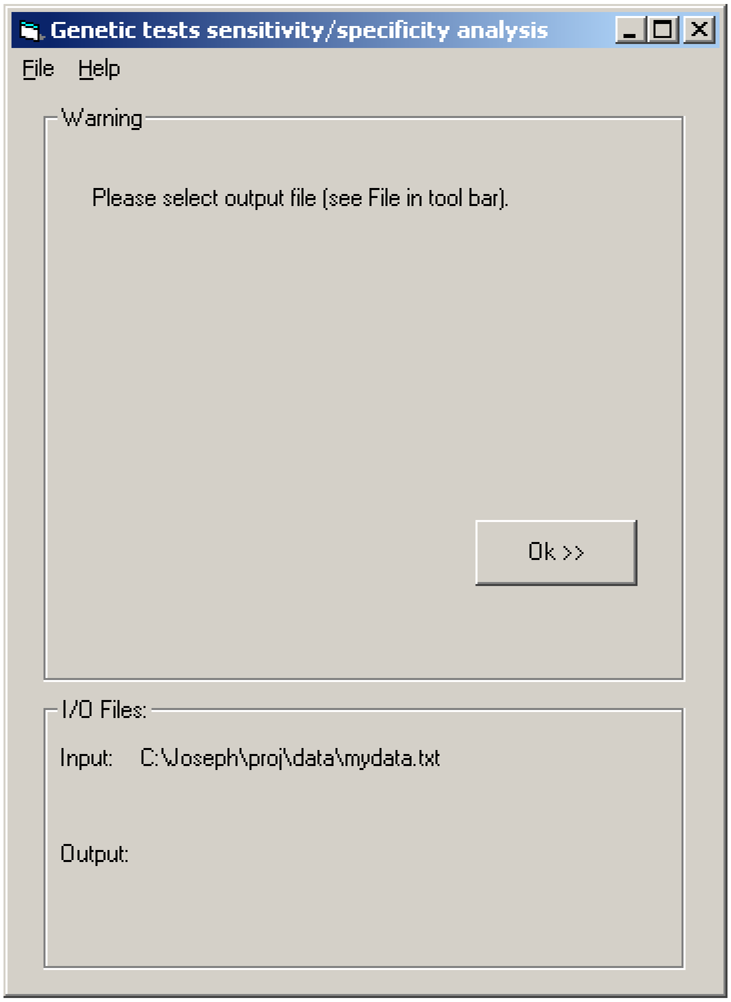


Subject # 14196 found twice in input file
C:\Joseph\proj\data\mydata.txt

Please review data before re-submitting.

Warnings

If you omit one of the points illustrated above before submitting, a warning message will be issued and program will not advance further unless you remedy to the situation.



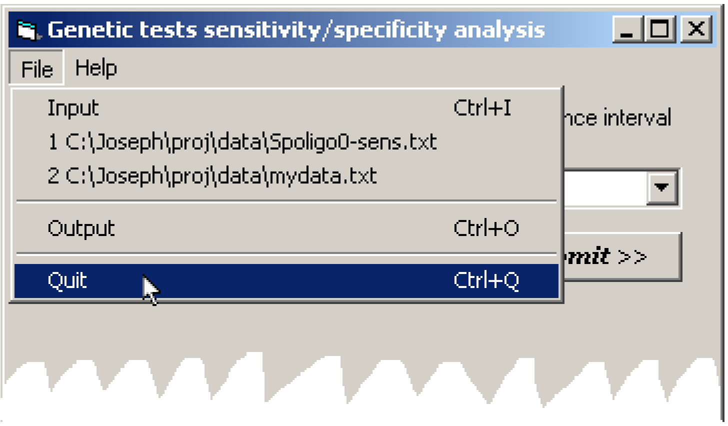
Please select output file (see File in tool bar).

Input: C:\Joseph\proj\data\mydata.txt

Output:

Quitting

When you are done, simply click File and Quit in the tool bar, or type Ctrl-Q as a shortcut.


Genetic tests sensitivity/specificity analysis
File
Help
Input Ctrl+I
1 C:\Joseph\proj\data\Spoligo0-sens.txt
2 C:\Joseph\proj\data\mydata.txt
Output Ctrl+O
Quit Ctrl+Q
nce interval
mit >>

If you experience any problem with this program, please do not hesitate to contact either
Lawrence Joseph lawrence.joseph@mcgill.ca)
or Patrick Bélisle (patrick.belisle@rimuhc.ca)

7. Sample Outputs

Specificity / Clustering

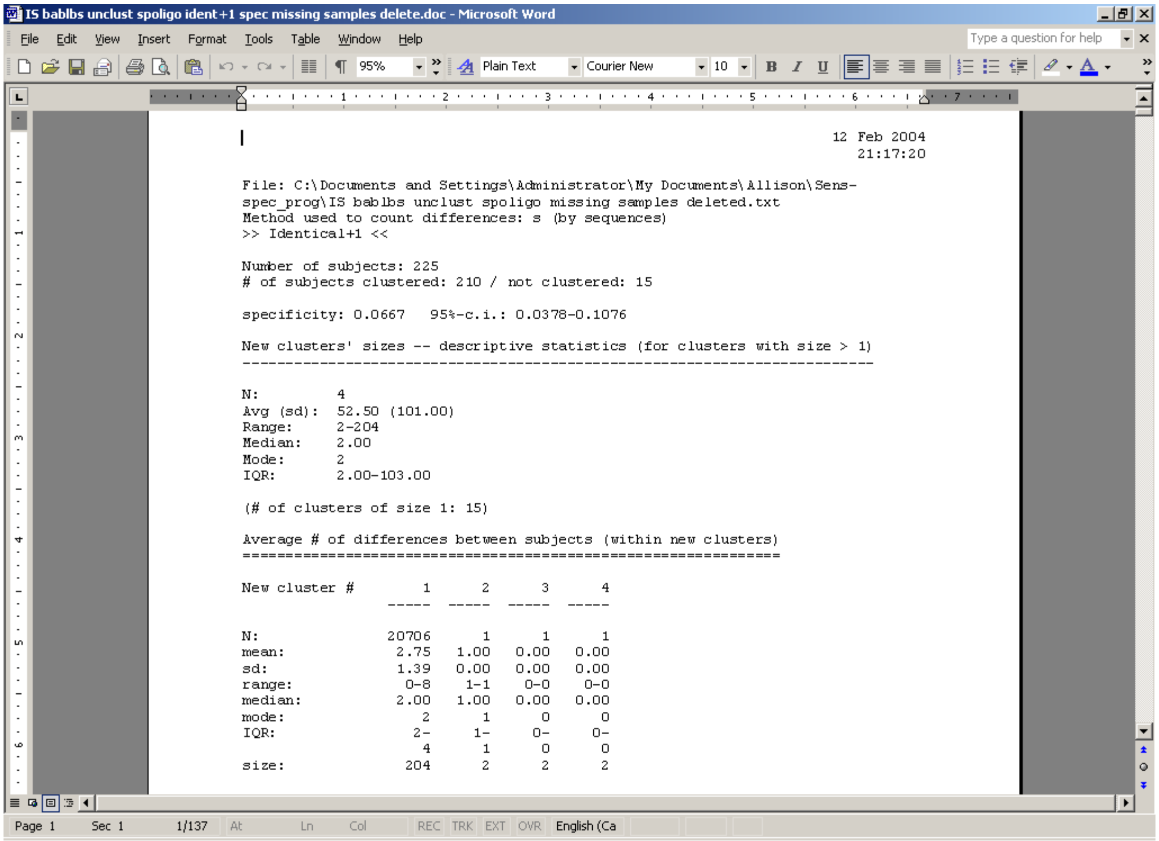
The figure that follows is a specificity output for spoligotyping, allowing clustered spoligotypes to have one difference, defined as one contiguous deletion. The descriptive statistics are as follows: The section labeled "New clusters' sizes – descriptive statistics" lists the number of spoligotype clusters, the mean and median number of isolates per cluster, as well as the range, interquartile range and mode.

The section entitled "Average # of differences between subjects (within new clusters)" lists, for each new cluster,

- the size of the cluster ("size")
- the number of pairwise comparisons within that cluster ("N")
- the mean and median number of differences between subjects (as well as the standard deviation, range, mode, and interquartile range)

If desired, an output can be generated that lists all the isolates and the number of differences between it and each other isolate.

Note: this is the output that would be generated if the program was being used just to calculate clusters. In that instance, the "specificity" calculated would actually be 1 – the % clustering.

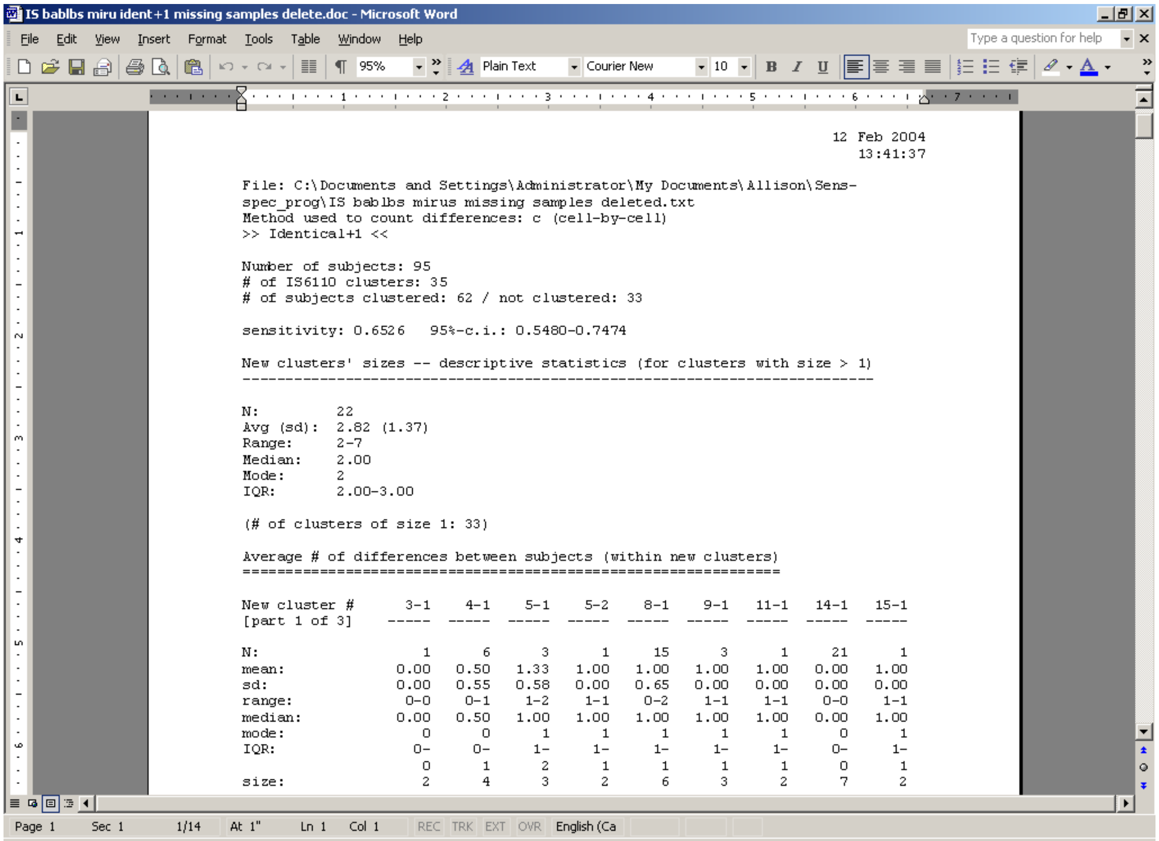

IS bablbs unclust spoligo ident+1 spec missing samples delete.doc - Microsoft Word
File Edit View Insert Format Tools Table Window Help
Type a question for help
95%
Plain Text
Courier New
10
Page 1 Sec 1 1/137 At Ln Col REC TRK EXT OVR English (Ca

```
                                                           12 Feb 2004
                                                              21:17:20

File: C:\Documents and Settings\Administrator\My Documents\Allison\Sens-
spec_prog\IS bablbs unclust spoligo missing samples deleted.txt
Method used to count differences: s (by sequences)
>> Identical+1 <<

Number of subjects: 225
# of subjects clustered: 210 / not clustered: 15

specificity: 0.0667   95%-c.i.: 0.0378-0.1076

New clusters' sizes -- descriptive statistics (for clusters with size > 1)
--------------------------------------------------------------------------

N:         4
Avg (sd):  52.50 (101.00)
Range:     2-204
Median:    2.00
Mode:      2
IQR:       2.00-103.00

 (# of clusters of size 1: 15)

Average # of differences between subjects (within new clusters)
===============================================================

New cluster #        1      2      3      4
                 -----  -----  -----  -----

N:               20706      1      1      1
mean:             2.75   1.00   0.00   0.00
sd:               1.39   0.00   0.00   0.00
range:             0-8    1-1    0-0    0-0
median:           2.00   1.00   0.00   0.00
mode:                2      1      0      0
IQR:                2-     1-     0-     0-
                     4      1      0      0
size:              204      2      2      2
```

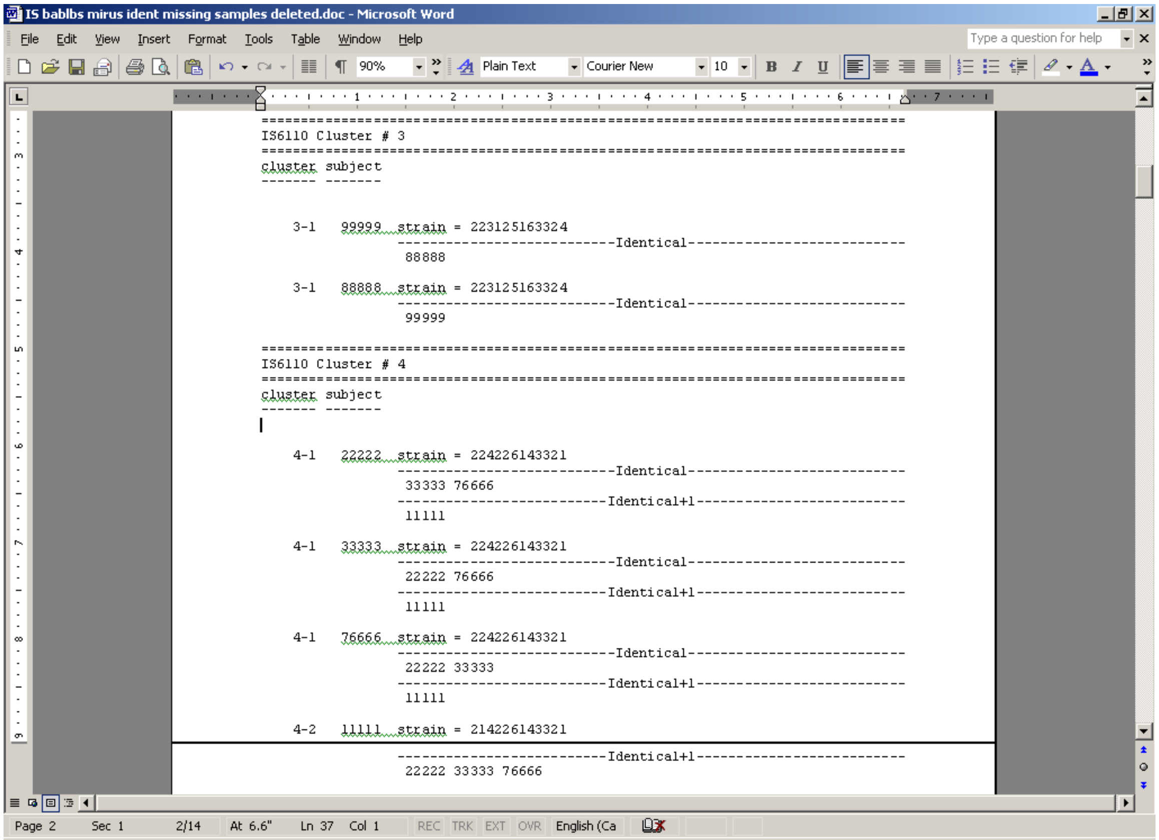
Sensitivity

The sensitivity output generated here is MIRU-VNTR, allowing one loci difference between clustered isolates. This are indicated by “cell-by-cell” and “ >> Identical +1 <<” on the output. The sensitivity output is similar to the specificity output, with a few additions:

- It lists the number of IS6110 RFLP clusters that were in the datafile (“# of IS6110 clusters”)
- Each new cluster has a two-part name: the IS6110 RFLP cluster number, then a dash and the MIRU-RFLP cluster number. For instance, IS6110 RFLP cluster #5 was broken down into two MIRU-VNTR clusters: cluster 5-1 and 5-2. IS6110 cluster #9 has only one MIRU-VNTR cluster, which has therefore been named 9-1.



```
                                                                12 Feb 2004
                                                                   13:41:37

File: C:\Documents and Settings\Administrator\My Documents\Allison\Sens-
spec_prog\IS bablbs mirus missing samples deleted.txt
Method used to count differences: c (cell-by-cell)
>> Identical+1 <<

Number of subjects: 95
# of IS6110 clusters: 35
# of subjects clustered: 62 / not clustered: 33

sensitivity: 0.6526   95%-c.i.: 0.5480-0.7474

New clusters' sizes -- descriptive statistics (for clusters with size > 1)
--------------------------------------------------------------------------


N:          22
Avg (sd):   2.82 (1.37)
Range:      2-7
Median:     2.00
Mode:       2
IQR:        2.00-3.00

(# of clusters of size 1: 33)

Average # of differences between subjects (within new clusters)
===============================================================


New cluster #      3-1    4-1    5-1    5-2    8-1    9-1   11-1   14-1   15-1
[part 1 of 3]    -----  -----  -----  -----  -----  -----  -----  -----  -----

N:                   1      6      3      1     15      3      1     21      1
mean:             0.00   0.50   1.33   1.00   1.00   1.00   1.00   0.00   1.00
sd:               0.00   0.55   0.58   0.00   0.65   0.00   0.00   0.00   0.00
range:             0-0    0-1    1-2    1-1    0-2    1-1    1-1    0-0    1-1
median:           0.00   0.50   1.00   1.00   1.00   1.00   1.00   0.00   1.00
mode:                0      0      1      1      1      1      1      0      1
IQR:                0-     0-     1-     1-     1-     1-     1-     0-     1-
                     0      1      2      1      1      1      1      0      1
size:                2      4      3      2      6      3      2      7      2
```



```
================================================================================
IS6110 Cluster # 3
================================================================================
cluster subject
------- -------


    3-1   99999  strain = 223125163324
                 ---------------------------Identical---------------------------
                  88888

    3-1   88888  strain = 223125163324
                 ---------------------------Identical---------------------------
                  99999

================================================================================
IS6110 Cluster # 4
================================================================================
cluster subject
------- -------


    4-1   22222  strain = 224226143321
                 ---------------------------Identical---------------------------
                  33333 76666
                 --------------------------Identical+1--------------------------
                  11111

    4-1   33333  strain = 224226143321
                 ---------------------------Identical---------------------------
                  22222 76666
                 --------------------------Identical+1--------------------------
                  11111

    4-1   76666  strain = 224226143321
                 ---------------------------Identical---------------------------
                  22222 33333
                 --------------------------Identical+1--------------------------
                  11111

    4-2   11111  strain = 214226143321
                 --------------------------Identical+1--------------------------
                  22222 33333 76666
```

References

Scott, A. N., D. Menzies, T. N. Tannenbaum, L. Thibert, R. Kozak, L. Joseph, K. Schwartzman, and M. A. Behr. 2005. Sensitivities and Specificities of Spoligotyping and Mycobacterial Interspersed Repetitive Unit–Variable-Number Tandem Repeat Typing Methods for Studying Molecular Epidemiology of Tuberculosis. J Clin.Microbiol. (In press)

Warren, R. M., E. M. Streicher, S. L. Sampson, G. D. Van Der Spuy, M. Richardson, D. Nguyen, M. A. Behr, T. C. Victor, and P. D. Van Helden. 2002. Microevolution of the direct repeat region of Mycobacterium tuberculosis: implications for interpretation of spoligotyping data. J.Clin.Microbiol. **40:**4457-4465.