DNA mixture separation using **mixsep**

Torben Tvedebrink
Department of Mathematical Sciences
Aalborg University
tvede@math.aau.dk

August 12, 2011

This vignette describes how the **mixsep** can be used to separate DNA mixtures of two and three DNA profiles. The implemented algorithm is based on the research paper by Tvedebrink et al. (2011). For more details and motivation please see that paper.

DISCLAIMER: *This vignette is not completely up-to-date compared to version of* ***mixsep*** *available at CRAN. It is expected to be fixed during July/August of 2011.*

Contents

1 Create a desktop shortcut (Windows only)

When R and **mixsep** are installed the `makeShortcut`-function in the **mixsep**-package can be used to create a shortcut on the Windows desktop.

There are several reasons to do so:

- Makes it easier and faster to start the **mixsep** GUI.
- The performance of tcl/tk within R is improved when it is run outside the RGui.
- If the `makeShortcut`-function is given a configuration file a database connection is readily available by start-up (see Section 2 for further details).

In Figure 1 a screen short of the RGui after the **mixsep**-package is loaded and the `makeShortcut`-function is executed. Note that in order to call `library(mixsep)` successfully the package needs to be available/installed in R. If no errors occur a shortcut icon to **mixsep** (similar to the one in Figure 1) is created on the Windows desktop.

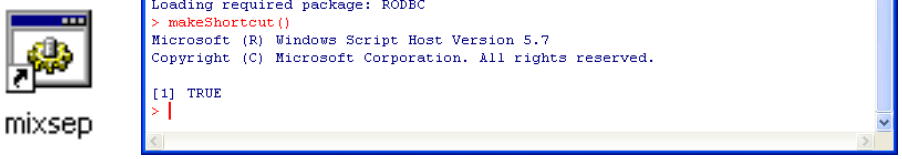


Figure 1: Shortcut creating under R for Windows

2 Database connection and data extraction

2.1 Setting up data connection to a database

Because **mixsep** depends on the **RODBC**-package it is possible to connect to a database and retrieve case information data without an intermediate create/read file step.

In order to do so one needs to give a connection file as argument to the `makeShortcut`-function (see the Section 1). One example of such a file is given below (the file name and extension is

unimportant). The **mixsep** GUI uses the `odbcDriverConnect`-function from the **RODBC**-package in order to connect to a database. Hence, please refer to the documentation for this function by typing `?odbcDriverConnect` in R or read the package Vignette by Brian Ripley.

The content of connection file (`tvede.connect`). It is important that only one argument (`db, dbtab, dbcase, dbcols`) is specified per line. Each argument is given in double-quotes (`"argument"`) and these are separated by semi-colon (`"argument1; argument2"`). Each line is described in more detail below:

```
## Settings for connecting to the MySQL server on Torben Tvedebrink's laptop
db = "DSN=mixsepServer; Database=mixsep"
dbtab = "samples"
dbcase = "cases"
dbcols = "replicate; fraction"
```

db The `db` argument gives the DSN connection name and specifications to `odbcDriverConnect`. In the example above the ODBC data source is called `mixsepServer` and the particular database `mixsep`.

dbtab The `dbtab` argument specifies which table on the `db` we access, e.g. a possible database query is `"SELECT * FROM dbtab"`.

dbcase The `dbcase` specifies the column containing the case number associated with a given case. Note that there may be several analysis results sharing `dbcase`-value if these are e.g. replicate runs, different PCR fractions, etc.

dbcols As mentioned under `dbcase` can several analysis result share `dbcase`-value. However, there should at least be one column making it possible to distinct between these. This/these columns are specified as `dbcols`.

The screen short in Figure 2 show the creation of a mixsep desktop shortcut while creating the database connection. **NOTE:** The connection file (here `tvede.connect`) needs to be in the *current working directory*, which is set by selecting "Change dir..." under "File" in the menu (see Figure 18). The chosen directory should contain the connection file.

2.2 Extract data from a database connection

By using the shortcut created above a database-button is available on the "Files"-tab of the **mixsep** GUI (Figure 3).

By clicking this button a "database query"-window opens in which queries to the specified database can be made (Figure 4). A query is made to the database matching all entries in the database with `dbcase` containing the specified case identifier. The result is stratified by `dbcols`.

In Figure 4 the second screen shot a query to the case *10-18687* is made and three different analysis results are returned (third screen shot of Figure 4). Furthermore, the call to the database uses *wild cards* such that any consecutive sub-string of *10-18687* would yield the three samples (and possibly more). Note that ":" is used to separate the column information from the columns specified in `dbcase` and `dbcols` above. Each row corresponds to an unique sample file (originating from different replicates and PCR runs).

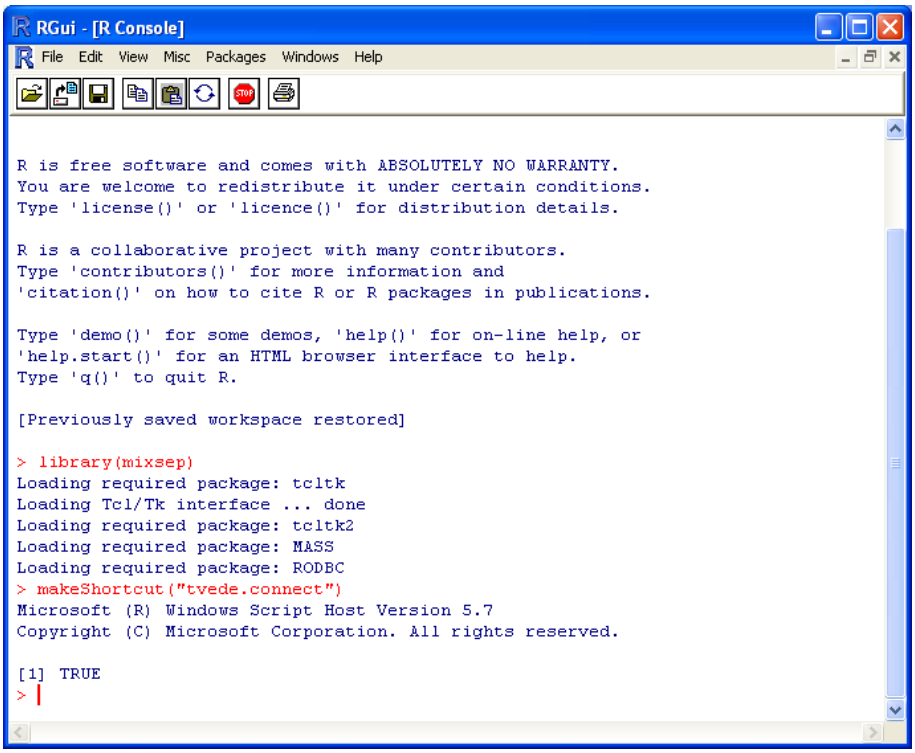


Figure 2: Creating a shortcut for the Windows desktop and setting up a database connection.

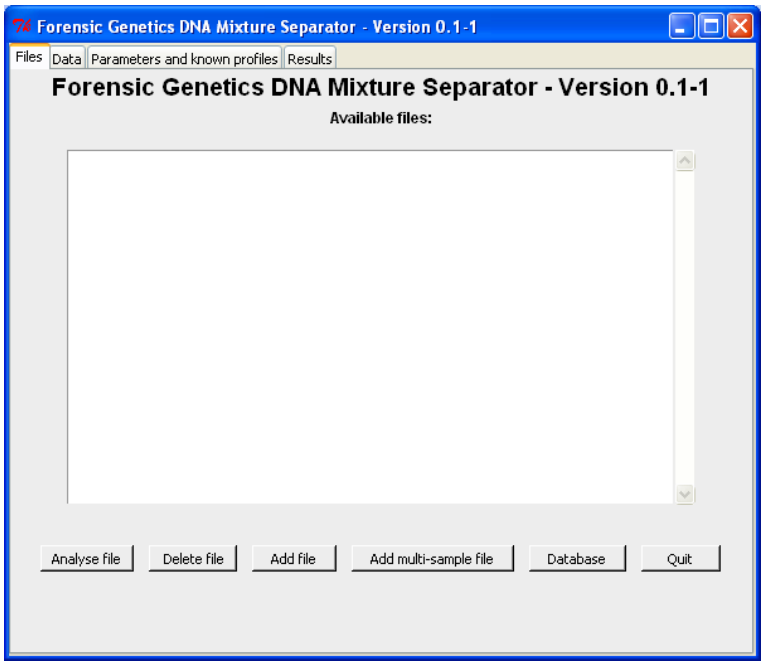
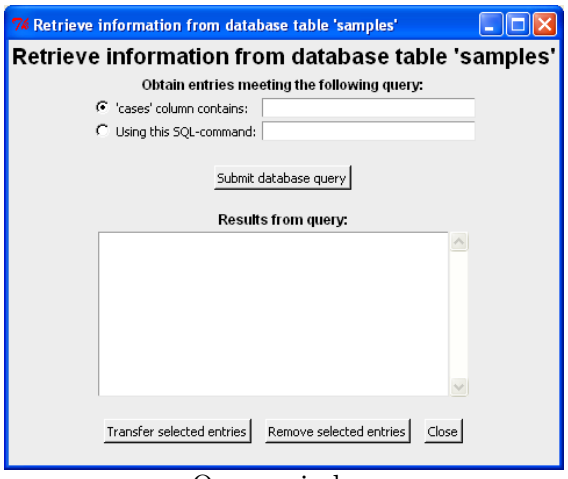


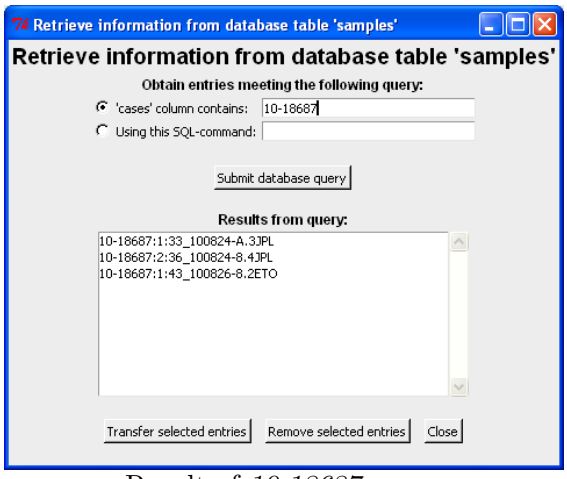
Figure 3: If a database connection is setup the Database-button appears on the "Files"-tab.

The sample(s) of interest can be transferred to the main window by double-click a the row or by highlighting the row(s) and click "Transfer selected entries". Note both [Shift], [Ctrl] and

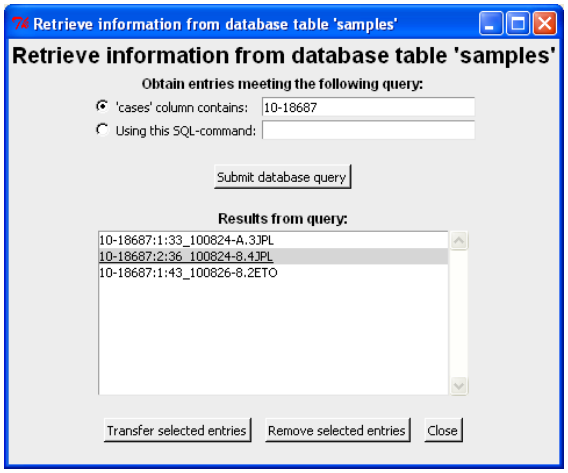
[Ctrl-a] can be used to select more/all samples. Here we select *10-18687:2:36_100824-8.4JPL* for further analysis (which is then transferred to the main window - see fourth screen shot of Figure 4):



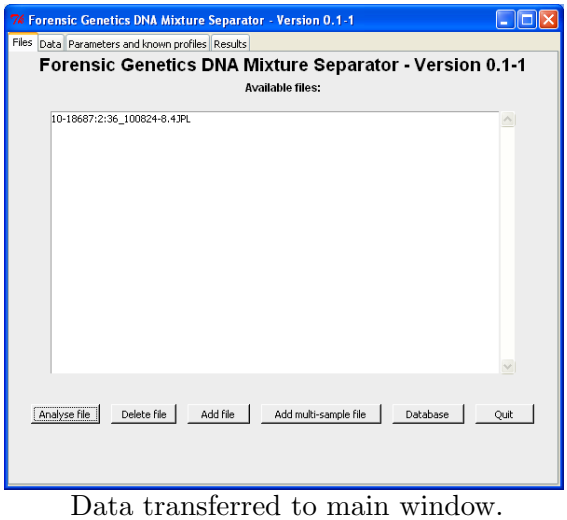
Query window.



Result of *10-18687* query.



Mark a entry for transfer.



Data transferred to main window.

Figure 4: Process of querying the database and transfer data to the main window.

3 Load data from files

3.1 Load data from file(s) with a single case and replicate

In order to load data from files these needs to be either .csv-files (semi-colon separated, *delimiter:* ";", or comma separated, *delimiter:* ",") or .tab-files (tabular separated). *Other formats may be included in later releases if requested by the users.*

To load file(s) with a single case (and replicate) use the "Add file"-button, which opens a "Open file"-dialogue window. One or more files can be selected (Figure 5).

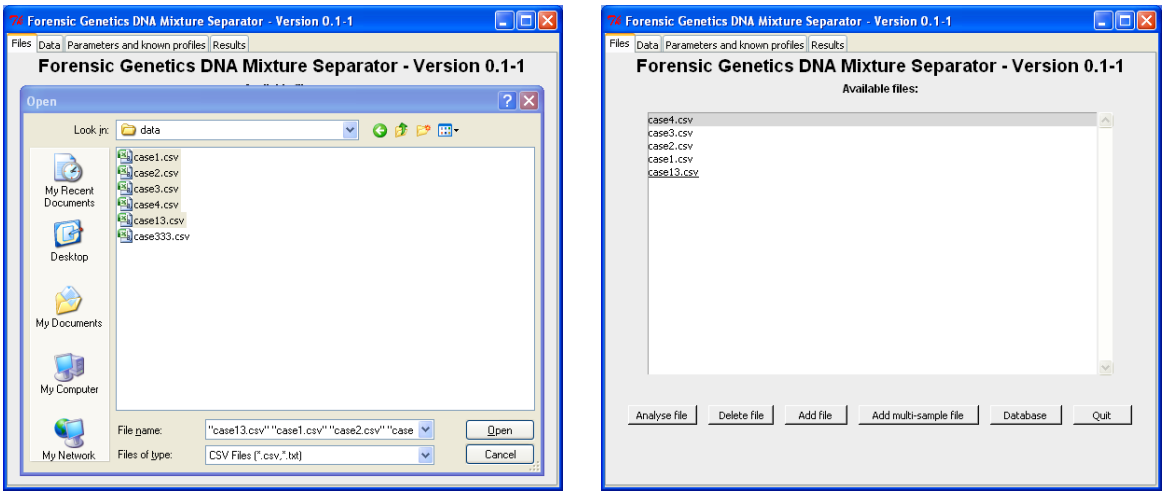


Figure 5: Opening several files and load of data into main window.

After clicking "Open" in the dialogue window, the selected file(s) are present in the "Available files" list for further analysis (Figure 5). In Section 4 it is demonstrated how a file is analysed.

3.2 Load data from file(s) with multiple cases/replicates

If a file contains more cases or replicates of the same case, the "Add multi-sample file"-button should be used. This opens a "Open file"-dialogue window, where one or more files can be selected (see single file description in Section 3.1). However, in order to specify which columns that uniquely determines the different samples a second dialogue window opens (top screen shot in Figure 6)

In the top screen shot of Figure 6 are *Proeve_ID* and *Fraction_Run_ID* selected since these uniquely separates the rows into three different samples. By clicking "Select sample column" the data is transferred to the main window. Note that *Proeve_ID* is fixed for all samples, but since this column contain the case number information we also mark this column (See bottom screen shot of Figure 6.

4 Determine the best matching configuration

We demonstrate how to determine the best matching configuration for a two-person DNA mixture using the wang.csv-file. This file contains data previously published by Wang et al. (2006). First we load the file (as described in Section 3.1) to get the file present in "Available files"-list as in Figure 7.

By marking the file name and click the "Analyse file"-button (or double-click on the file name), we obtain the "Data"-tab shown in Figure 8. In the "Data"-tab the columns containing the relevant information are selected (left screen shot). Since wang.csv only contain information about locus, allele and area, we leave "Height", "bp" and "Dye" empty.

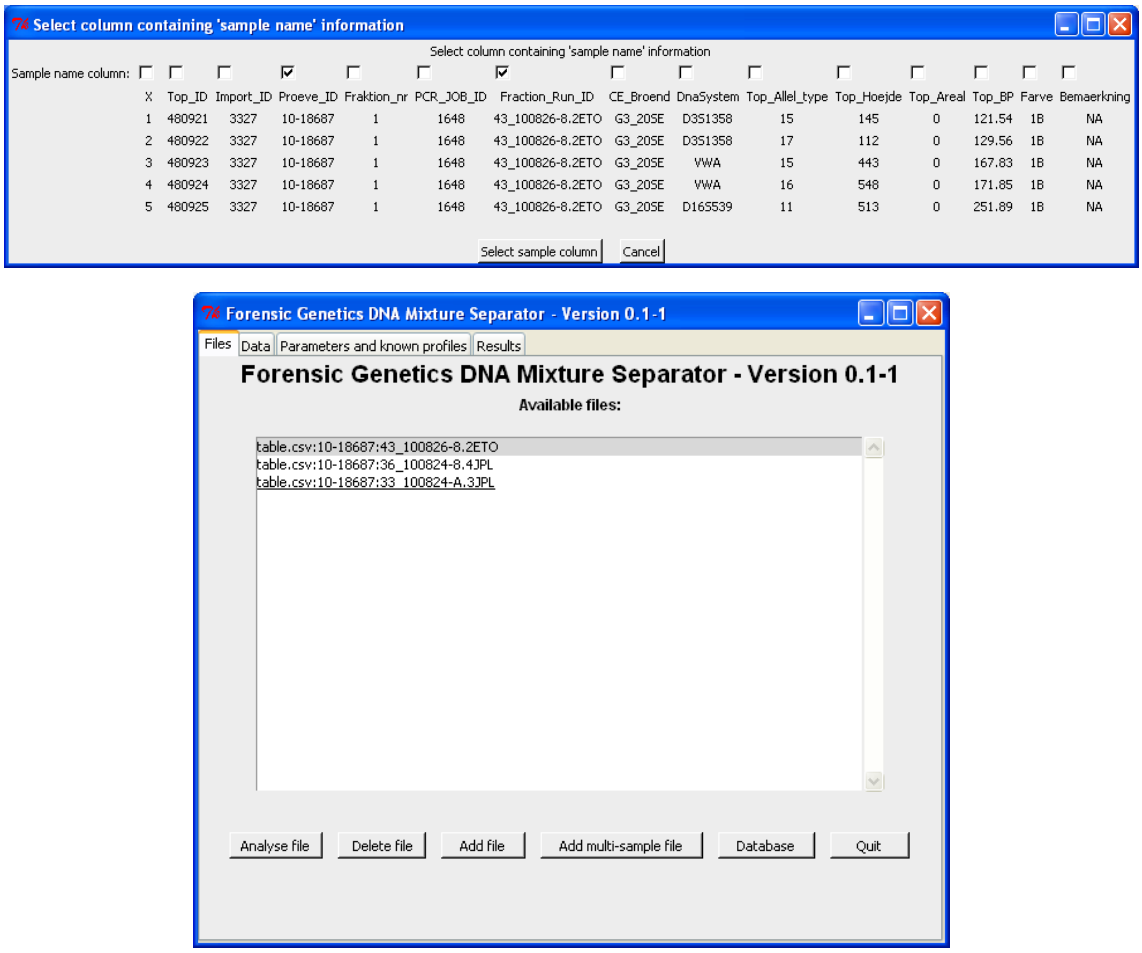


Figure 6: Multi-sample files loaded into the main window.

After the columns are selected (by clicking "Select columns"), we obtain all the rows (for the selected columns) in the file (right screen shot of Figure 8). For each row it is possible to select/remove the peak from the analysis. However, we recommend that the removal of rows is used with caution, since all information should be considered and dealt with!

The "Parameters and known profiles"-tab makes it possible to specify the number of contributors, whether the algorithm should search for alternatives, etc. It is also here that fixed/known profiles can be specified (see Section 5 for an analysis with fixed profiles). From the parameter settings in Figure 9, we assume the sample is a two-person mixture and searches for alternatives (the lower the level of significance is, the more alternatives are included in the final result).

When clicking "Analyse mixture!" the computer resolves the DNA mixture and returns result in the "Result"-tab. Figure 10 has the output from the analysis. For each locus a best matching configuration (which is identical to the true profiles, see by Wang et al. (2006)) is identified together with a list of possible alternatives. Next to the locus designation is the number of possible alternatives for that locus given in parenthesis. Below the list of locus configurations is the number of total configurations given together with the estimated

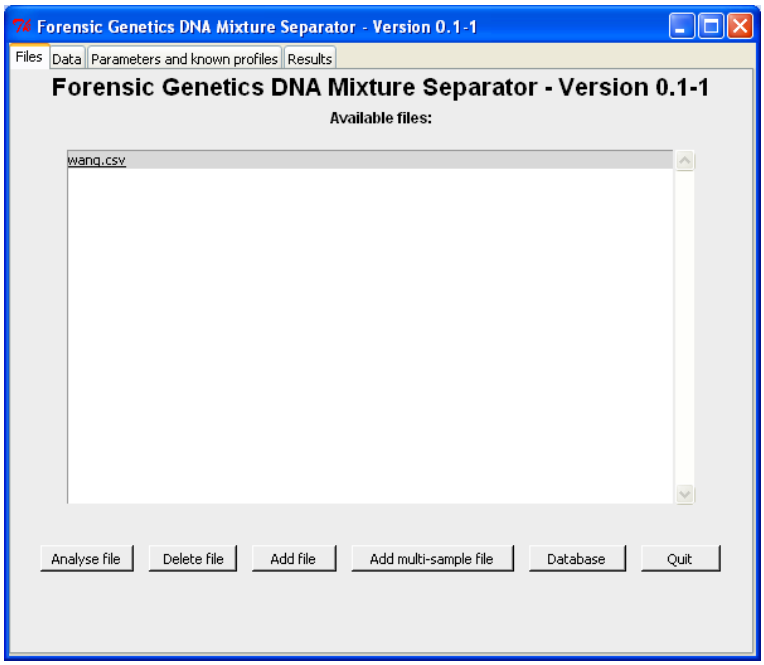


Figure 7: The tutorial data file "wang.csv" is loaded.

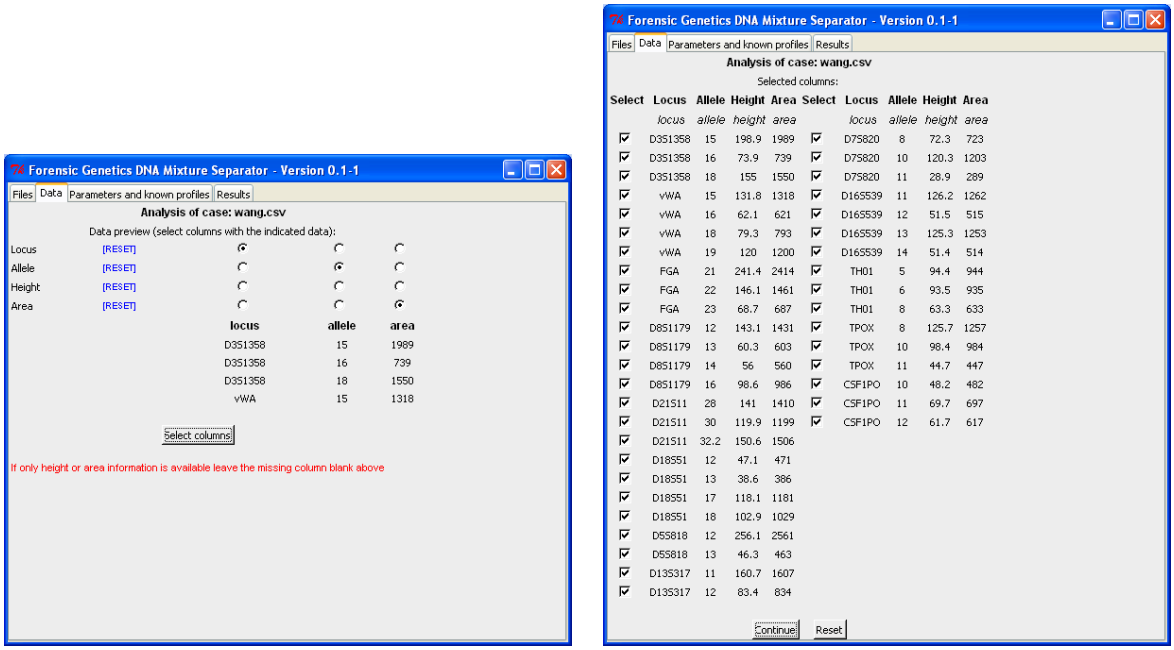


Figure 8: The two states of the "data"-tab. Left: Select the relevant columns. Right: Select the rows of interest (all rows are selected by default).

parameters of the statistical model. The "Estimated alpha" represents the mixture proportion of the minor contributor. Here it is 0.3 which corresponds to an 1:2-mixture ratio. The "Estimated tau" value represents that residual variance, which means that the smaller the

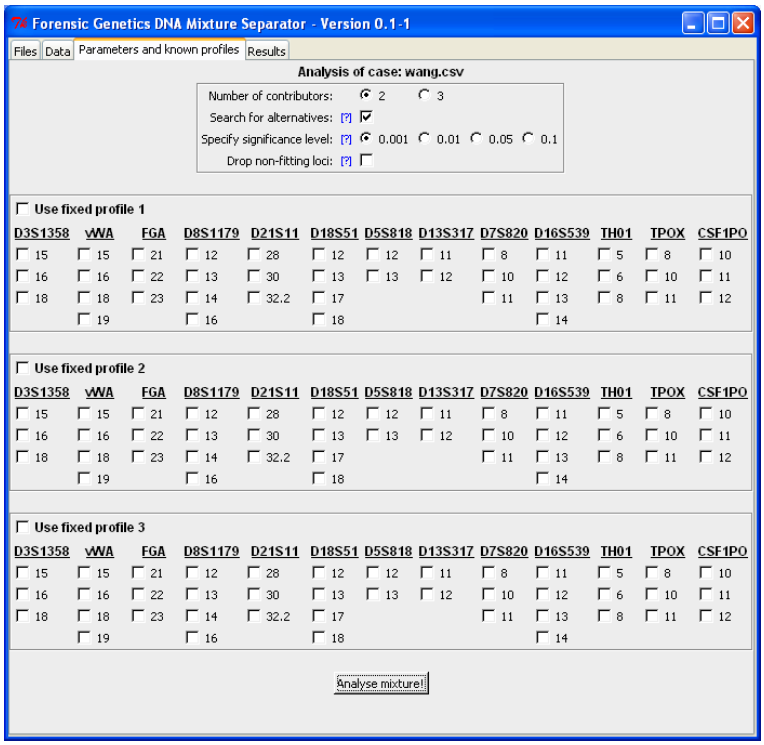


Figure 9: Parameter settings implying a two-person DNA mixture and search for alternative configurations.

estimate the better is the concordance between the observed and expected peak intensities (the "Derived R^2" quantity is explained below).

The expected peak intensities for the selected configuration can be plotted against the observed peak intensities in a EPG-like plot (Figure 11). The coloured cones represents the observed peak intensities, while the black lined cones show the expected peak intensities (for this particular plot it is hard to see any discrepancies, which indicate a good fit between the observed and expected peak intensities).

From the analysis output alternative configurations can be specified by marking the combinations in the lists. For each locus the list of alternatives is sorted in decreasing order in terms of goodness-of-fit. For the selected configuration (Figure 12) the R^2-quantity is computed as the ratio of the estimated tau values. Here $R\hat{}2 = 0.38 = 118/310$ and the closer R^2 is to 1, the better is the alternative configuration relative to the best match pair of profiles.

In the EPG-plot of Figure 13 we see a change in the fit between the black lined cones and coloured cones. In addition to the plot it is possible to export the results to a ASCII text file by clicking "Export results". For this example the text file is printed on page 9.

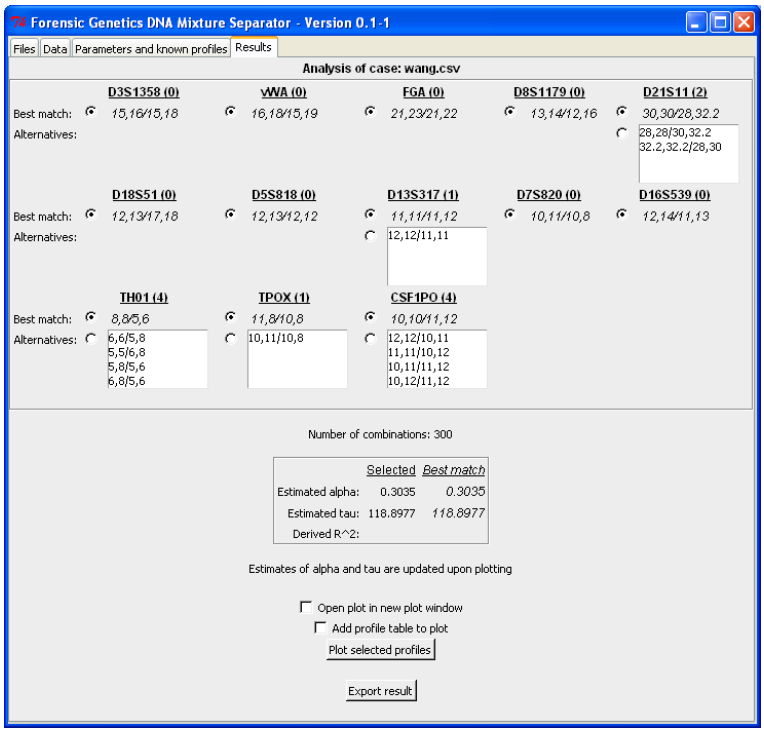


Figure 10: Results from the best matching analysis.

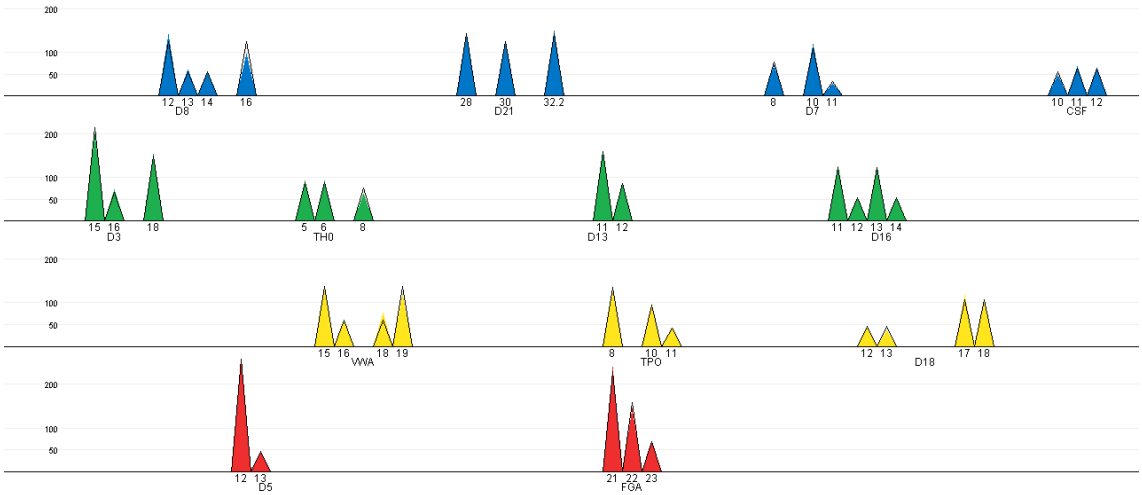


Figure 11: EPG for the best matching pair of DNA profiles. The coloured cones show the observed peak intensities whereas the solid black cones represents the expected peak intensities.

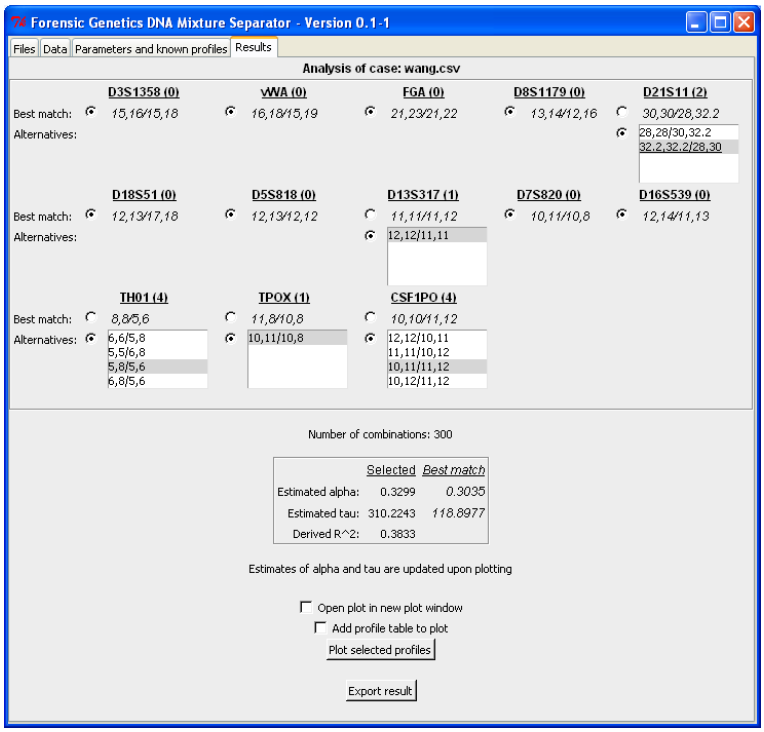


Figure 12: Selecting an alternative pair of profiles from the list of locus-wise alternatives.

```
MixSep Output

Analysis of case file: wang.csv

==== PROFILES ====

             D3S1358        vWA            FGA            D8S1179
Best match   15,16/15,18*   16,18/15,19*   21,23/21,22*   13,14/12,16*
Alternatives


             D21S11             D18S51         D5S818         D13S317
Best match   30,30/28,32.2      12,13/17,18*   12,13/12,12*   11,11/11,12
Alternatives 28,28/30,32.2                                   12,12/11,11*
             32.2,32.2/28,30*


             D7S820         D16S539        TH01       TPOX          CSF1PO
Best match   10,11/10,8*    12,14/11,13*   8,8/5,6    11,8/10,8     10,10/11,12
Alternatives                               6,6/5,8    10,11/10,8*   12,12/10,11
                                           5,5/6,8                  11,11/10,12
                                           5,8/5,6*                 10,11/11,12*
                                           6,8/5,6                  10,12/11,12

=== PARAMETERS ===
```

```
             alpha  tau      R2
Best match   0.3035 118.8977
Selected (*) 0.3299 310.2243 0.3833


==== SETTINGS ====

Number of contributors: 2
Level of significance: 0.001
Number of combinations: 300
```

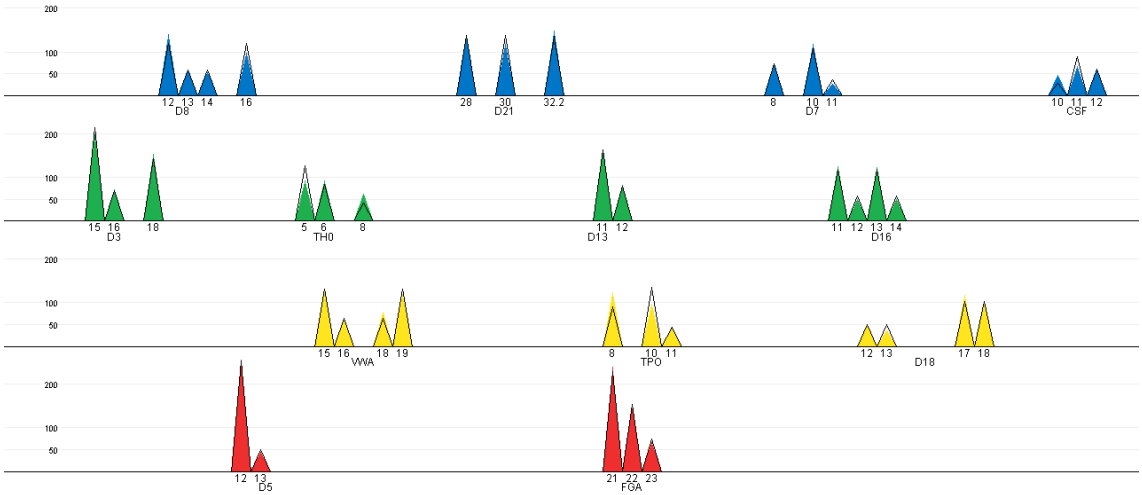


Figure 13: EPG of the alternative configuration specified in Figure 12.

5 Analysis with a fixed profile

For the initial steps of loading data files and selecting columns/rows for further analysis, please see Section 4 for a guide for determining the best matching configuration). In this guide we demonstrate how to analyse the same sample using a fixed (e.g. suspect) profile.

The "Parameters and known profiles"-tab makes it possible to specify the number of contributors, whether the algorithm should search for alternatives and fixed/known profiles. We assume the sample is a two-person mixture and searches for alternatives (the lower the level of significance is, the more alternatives are included in the final result). Furthermore, we want to investigate whether the profile in Table 1 a possible contributor to the DNA mixture (see parameter settings in the left screen shot of Figure 14).

When clicking "Analyse mixture!" the computer resolves the DNA mixture and returns result in the "Result"-tab. In the right screen shot of Figure 14 is the output from the analysis. For each locus the unknown profile that together with the fixed profile best explains the DNA mixture is identified (denoted F1/U for Fixed profile 1 and Unknown). In the U/U-row is the best matching configurations (both profiles unspecified, i.e. U/U is short

Table 1: DNA profiles of the true offender and the suspect used in the example. The bold font alleles in the suspect's profile denote the difference from the true offender profile (the major profile).

Locus	D3	vWA	FGA	D8	D21	D18	D5	D13	D7	D16	TH01	TPOX	CFS
Offender	15,18	15,19	21,22	12,16	28,32.2	17,18	12,12	11,12	8,10	11,13	5,6	8,10	11,12
Suspect	15,18	15,19	21,22	12,16	28,32.2	17,18	12,12	11,**11**	8,10	11,13	5,**8**	8,10	**12**,12

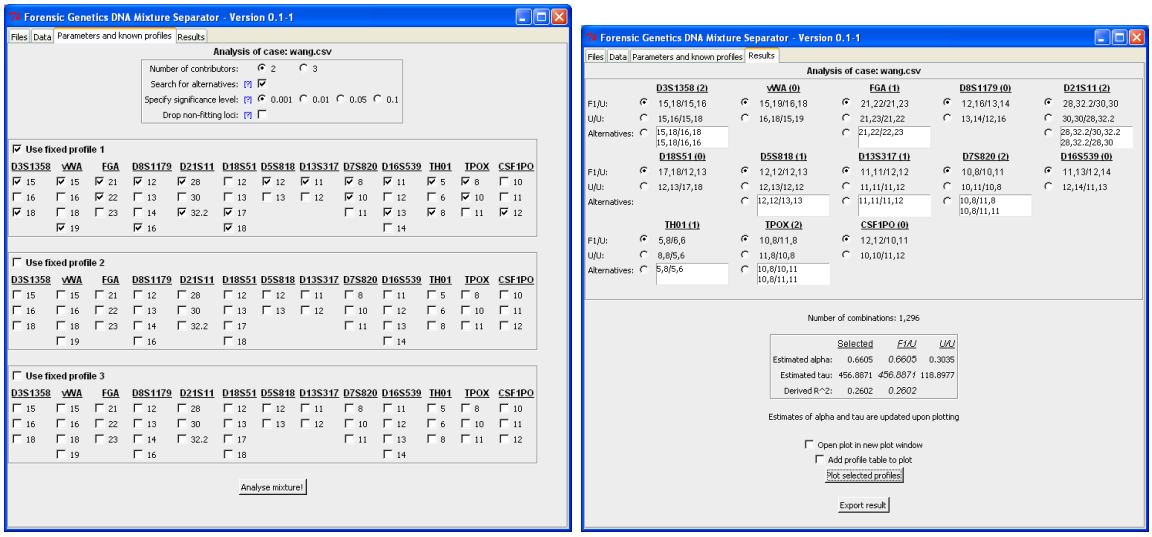


Figure 14: Parameter settings assuming a two-person DNA mixture with a known/fixed contributor as specified by the tick-boxes.

for Unknown/Unknown). Furthermore, below is a list of possible alternative configurations including the fixed profile.

Next to the locus designation is the number of possible alternatives for that locus given in parenthesis. Below the list of locus configurations is the number of total configurations given together with the estimated parameters of the statistical model. The "Estimated alpha" for F1/U represents the mixture proportion of the fixed contributor, whereas for U/U it denotes the contribution from the minor contributor. Here alpha is respectively 0.66 and 0.3, which approximately corresponds to an 1:2-mixture ratio in both cases. The "Estimated tau" value represents that residual variance, which means that the smaller the estimate the better is the concordance between the observed and expected peak intensities. The "Derived R^2" is computed as the tau for U/U to tau for F1/U. Here R^2 = 0.26 = 118/457 implying that the residual variance is four times bigger when the suspect's profile is assumed in the mixture compared to the best matching pair (which is identical to the true contributors, see Section 4).

The expected peak intensities for the selected configuration can be plotted against the observed peak intensities in a EPG-like plot. The coloured cones represents the observed peak intensities, while the black lined cones show the expected peak intensities (Figure 15). Note that the configuration on CSF causes the expected peak intensities to deviate from the observed. This is also the only combination of the three changes between the true offender and suspect (see Table 1), that is not among the alternatives in the best matching analysis (see

the results of best matching analysis in Section 4 - where it should be remembered that the suspect corresponds to the major profile).

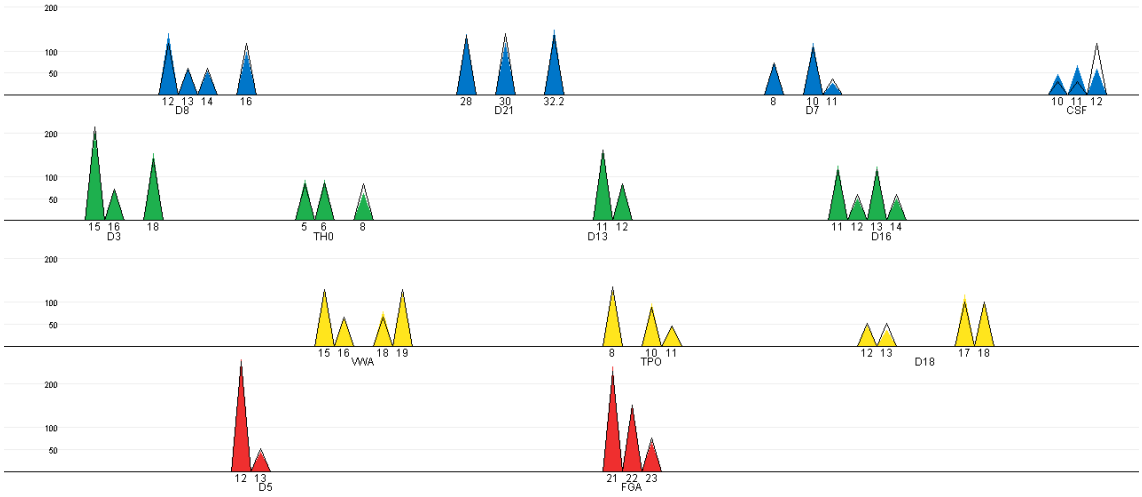


Figure 15: EPG with the suspect and best matching unknown DNA profile.

From the analysis output alternative configurations can be specified by marking the combinations in the lists. For each locus the list of alternatives is sorted in decreasing order in terms of goodness-of-fit. For the selected configuration the R^2 quantity is computed as the ratio of the estimated tau values of the best matching (U/U) and the selected configuration. In Figure 16 the R^2 = 0.06 = 118/2055 and the closer R^2 is to 1, the better is the alternative configuration relative to the best match pair of profiles.

In the EPG-plot in Figure 17 we see a change in the fit between the black lined cones and coloured cones. The low value of R^2 is indicates a poor fit between the observed and expected peak intensities, which is pictured in Figure 17. In addition to the plot it is possible to export the results to a ASCII text file by clicking "Export results" as in Section 4.

A Screen shots of basic R procedures

References

T. Wang, N. Xue, J. D. Birdwell (2006). Least-Square Deconvolution: A Framework for Interpreting Short Tandem Repeat Mixtures. Journal of Forensic Science 51 (6) (2006) 1284–1297. DOI: 10.1111/j.1556-4029.2006.00268.x

T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling (2011). Identifying contributors of DNA mixtures by means of quantitative information of STR typing. Computational Biology (In Press). Link: http://www.liebertonline.com/doi/abs/10.1089/cmb.2010.0055

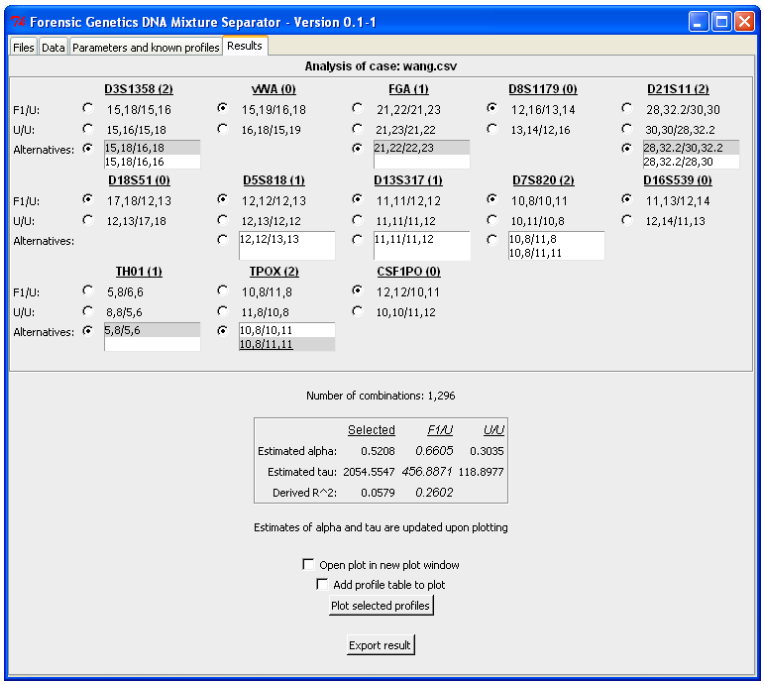


Figure 16: Specifying a different configuration (still involving the suspect).

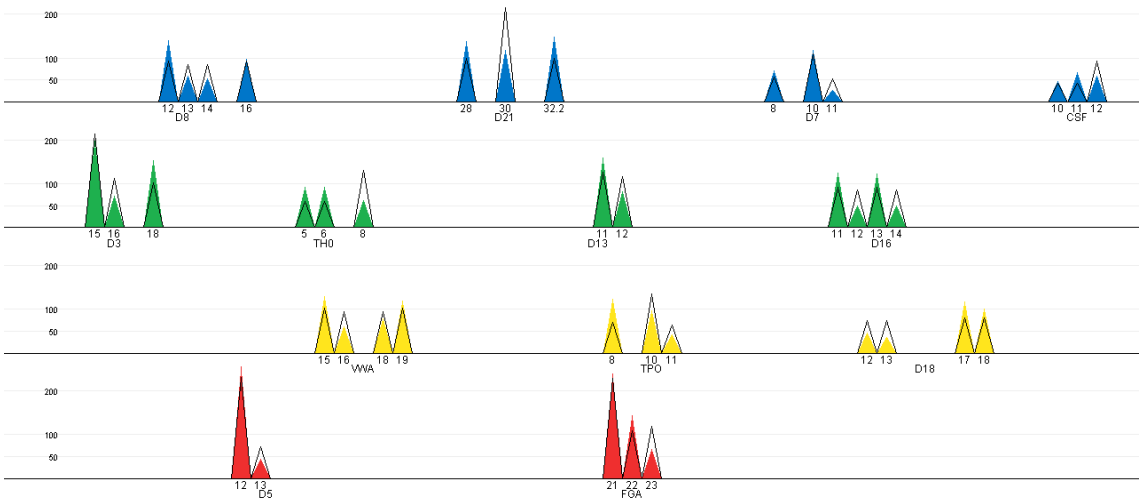


Figure 17: EPG of the configuration specified in Figure 16.

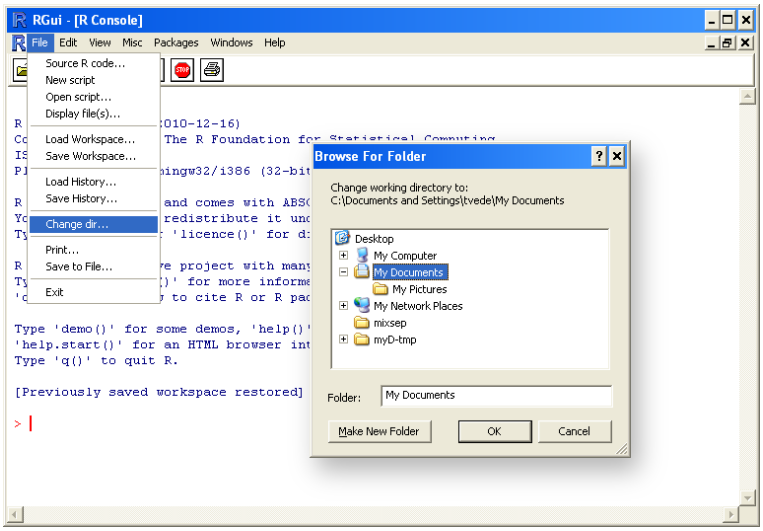


Figure 18: Changing the working directory in the R Gui on Windows.