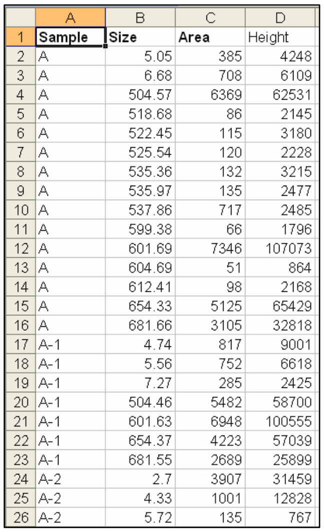
interactive_binner.r[1] manual

Alban Ramette
October 13, 2008

1. Preparing the input file

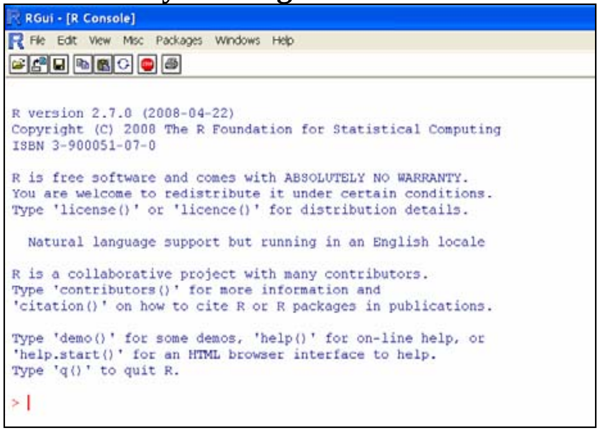
The GeneMapper output file containing the peak sizes, area and height can be copied to your favorite tabulation software.



	A	B	C	D
1	Sample	Size	Area	Height
2	A	5.05	385	4248
3	A	6.68	708	6109
4	A	504.57	6369	62531
5	A	518.68	86	2145
6	A	522.45	115	3180
7	A	525.54	120	2228
8	A	535.36	132	3215
9	A	535.97	135	2477
10	A	537.86	717	2485
11	A	599.38	66	1796
12	A	601.69	7346	107073
13	A	604.69	51	864
14	A	612.41	98	2168
15	A	654.33	5125	65429
16	A	681.66	3105	32818
17	A-1	4.74	817	9001
18	A-1	5.56	752	6618
19	A-1	7.27	285	2425
20	A-1	504.46	5482	58700
21	A-1	601.63	6948	100555
22	A-1	654.37	4223	57039
23	A-1	681.55	2689	25899
24	A-2	2.7	3907	31459
25	A-2	4.33	1001	12828
26	A-2	5.72	135	767

Copy the sample, size and area columns **only** to a text file (the height column is not needed). It is important to remove the lines that contain missing information. Column labels must be indicated. An example is given in Data for binner.xls in the "initial" sheet and in the corresponding GeneMapperData1.txt.

2. Start R by clicking on the R icon



```
R version 2.7.0 (2008-04-22)
Copyright (C) 2008 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

---

[1] This program is a free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.
This program is distributed in the hope that it will be useful, but **without any warranty**; without even the implied warranty of **merchantability** or **fitness for a particular purpose**. See the GNU General Public License for more details (Free Software Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA 02111-1307, USA)

3. Load the data into the R workspace

At the prompt (>), indicate in which directory you want to read and write the data (i.e. where you also put your .txt file. The directory should be created beforehand), and press enter. Note that quotes and \\ are used to indicate the path to the directory.

```
>setwd("c:\\R\\ARISA")
```

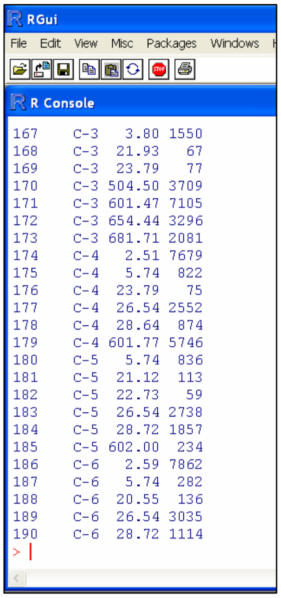
Then, load the data into the object D by typing the following: (make sure to exactly type the dots, commas, and punctuation signs, as indicated and use " instead of ”)

```
> D1=read.table("GeneMapperData1.txt",h=TRUE)
```

If you now type

```
>D1
```

You should see your data table appearing in the R console (if it is a big table, you will not see the first rows, but just the end of the table):

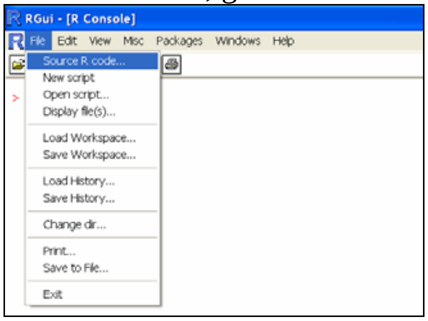


```
167    C-3   3.80 1550
168    C-3  21.93   67
169    C-3  23.79   77
170    C-3 504.50 3709
171    C-3 601.47 7105
172    C-3 654.44 3296
173    C-3 681.71 2081
174    C-4   2.51 7679
175    C-4   5.74  822
176    C-4  23.79   75
177    C-4  26.54 2552
178    C-4  28.64  874
179    C-4 601.77 5746
180    C-5   5.74  836
181    C-5  21.12  113
182    C-5  22.73   59
183    C-5  26.54 2738
184    C-5  28.72 1857
185    C-5 602.00  234
186    C-6   2.59 7862
187    C-6   5.74  282
188    C-6  20.55  136
189    C-6  26.54 3035
190    C-6  28.72 1114
> 
```

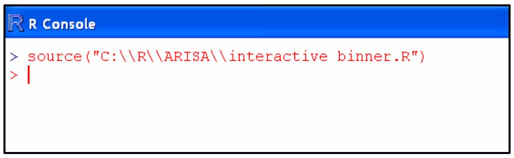
We are now ready to run the R script on the data stored in `D1`.

4. Running the interactive binner script

In the menu bar, go to Source R code…

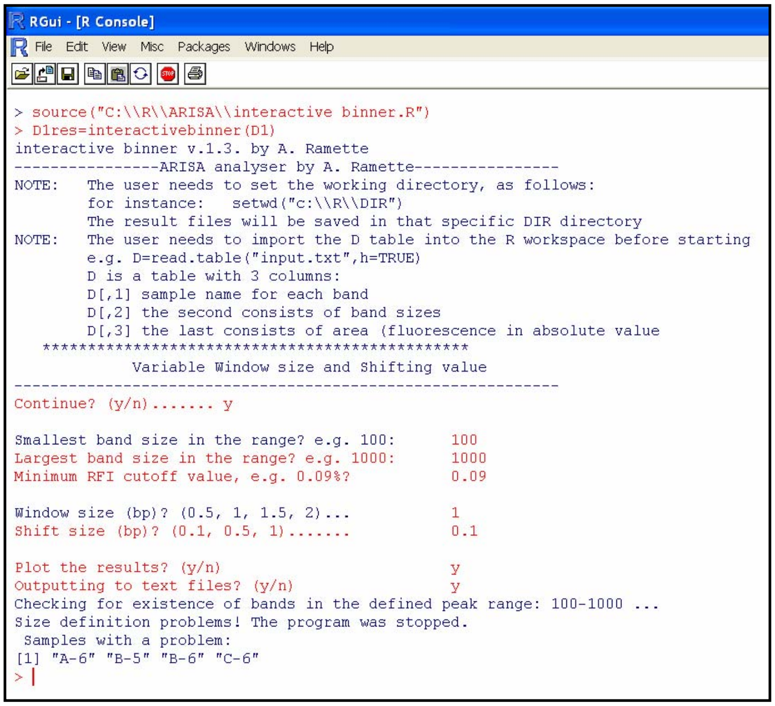


And indicate the location of your saved version of the interactive_binner.r script.



```
> source("C:\\R\\ARISA\\interactive binner.R")
>
```

then type the function name to be applied to `D1`:
>`D1res=interactivebinner(D1)`
(this command applies the function to `D1` and stores the results in `D1res`)



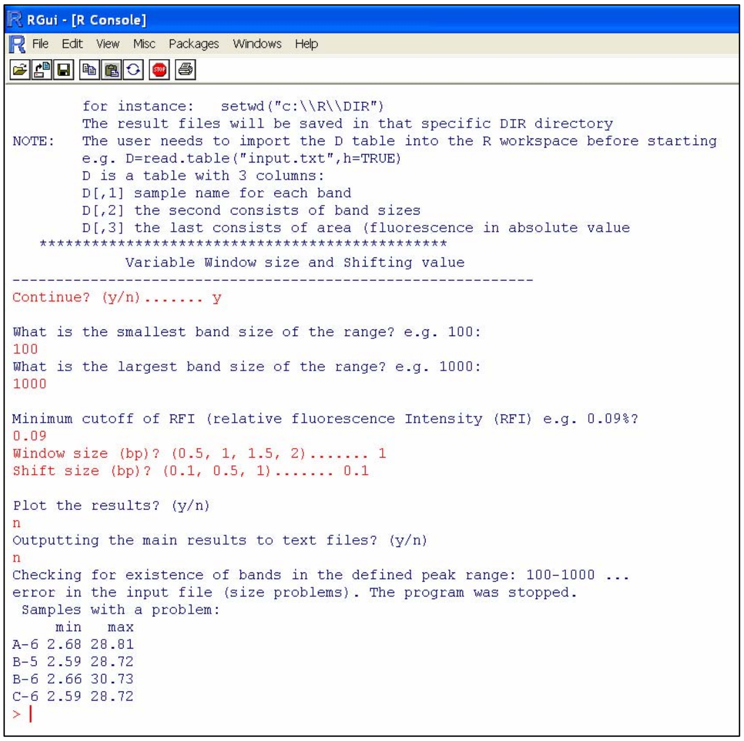
```
> source("C:\\R\\ARISA\\interactive binner.R")
> D1res=interactivebinner(D1)
interactive binner v.1.3. by A. Ramette
----------------ARISA analyser by A. Ramette----------------
NOTE:   The user needs to set the working directory, as follows:
        for instance:   setwd("c:\\R\\DIR")
        The result files will be saved in that specific DIR directory
NOTE:   The user needs to import the D table into the R workspace before starting
        e.g. D=read.table("input.txt",h=TRUE)
        D is a table with 3 columns:
        D[,1] sample name for each band
        D[,2] the second consists of band sizes
        D[,3] the last consists of area (fluorescence in absolute value
   ***********************************************
             Variable Window size and Shifting value
------------------------------------------------------------
Continue? (y/n)....... y

Smallest band size in the range? e.g. 100:      100
Largest band size in the range? e.g. 1000:      1000
Minimum RFI cutoff value, e.g. 0.09%?           0.09

Window size (bp)? (0.5, 1, 1.5, 2)...           1
Shift size (bp)? (0.1, 0.5, 1).......           0.1

Plot the results? (y/n)                         y
Outputting to text files? (y/n)                 y
Checking for existence of bands in the defined peak range: 100-1000 ...
Size definition problems! The program was stopped.
 Samples with a problem:
[1] "A-6" "B-5" "B-6" "C-6"
>
```

The script starts by indicating some basic information about the version, expected data format and ask you if you want to proceed. Type “`y`” (without the quotes), if the data table corresponds to the description provided, otherwise type “`n`” and see the points above.

For this example, we can use the following parameters:
- Smallest band size of the range `100`
- Largest band size of the range `1000`
- Minimum RFI cutoff of RFI `0.09`
- Window size `1`
- Shift size `0.1`

(type “`y`” for both plotting and outputting).

After few seconds, you should see the following message in the R console, indicating that the calculations are done:



```
        for instance:   setwd("c:\\R\\DIR")
        The result files will be saved in that specific DIR directory
NOTE:   The user needs to import the D table into the R workspace before starting
        e.g. D=read.table("input.txt",h=TRUE)
        D is a table with 3 columns:
        D[,1] sample name for each band
        D[,2] the second consists of band sizes
        D[,3] the last consists of area (fluorescence in absolute value
   ***********************************************
             Variable Window size and Shifting value
------------------------------------------------------------
Continue? (y/n)....... y

What is the smallest band size of the range? e.g. 100:
100
What is the largest band size of the range? e.g. 1000:
1000

Minimum cutoff of RFI (relative fluorescence Intensity (RFI) e.g. 0.09%?
0.09
Window size (bp)? (0.5, 1, 1.5, 2)....... 1
Shift size (bp)? (0.1, 0.5, 1)....... 0.1

Plot the results? (y/n)
n
Outputting the main results to text files? (y/n)
n
Checking for existence of bands in the defined peak range: 100-1000 ...
error in the input file (size problems). The program was stopped.
 Samples with a problem:
     min   max
A-6 2.68 28.81
B-5 2.59 28.72
B-6 2.66 30.73
C-6 2.59 28.72
>
```

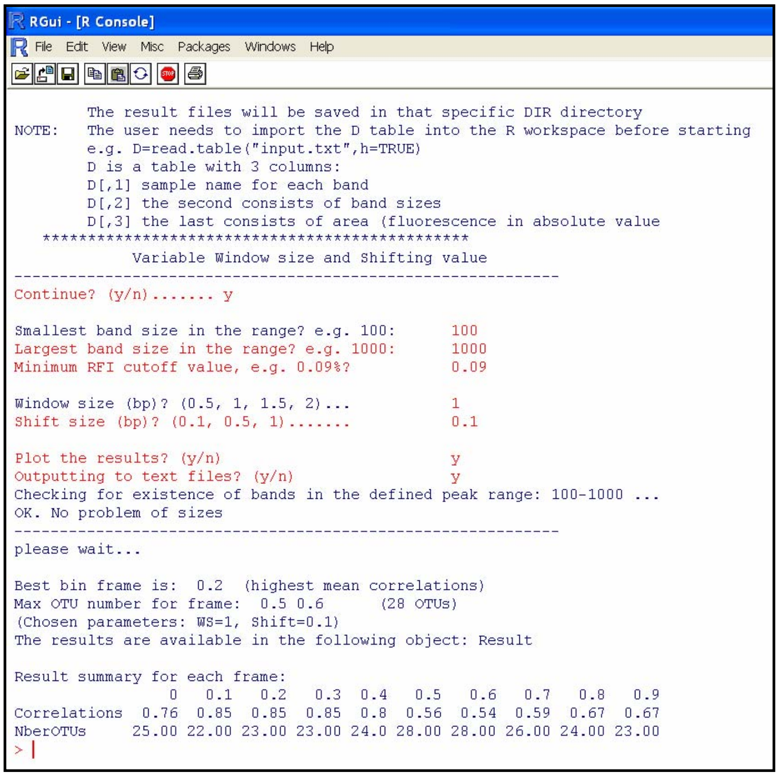
In this example, the script detected that for some samples, the highest peak size was not fitting in the predefined size range (e.g. for A-6 the largest peak was 28.8 bp while the selected range was 100-1000 bp). Those samples must then be manually removed from the data (i.e. you need to go back to point 1).

In this tutorial, the corrected files are found in Data for binner.xls (“corrected” sheet) and the corresponding GeneMapperData2.txt.

Reimport the data to the R workspace:

```
>D2=read.table("GeneMapperData2.txt",h=TRUE)
> D2res=interactivebinner(D2)
```

And run the script again. You should now obtain the following output.



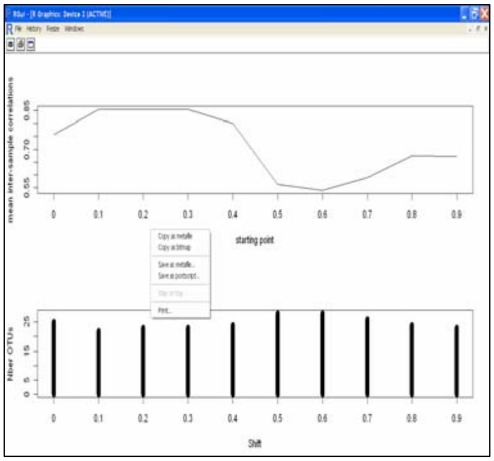
```
        The result files will be saved in that specific DIR directory
NOTE:   The user needs to import the D table into the R workspace before starting
        e.g. D=read.table("input.txt",h=TRUE)
        D is a table with 3 columns:
        D[,1] sample name for each band
        D[,2] the second consists of band sizes
        D[,3] the last consists of area (fluorescence in absolute value
   ***********************************************
             Variable Window size and Shifting value
------------------------------------------------------------
Continue? (y/n)....... y

Smallest band size in the range? e.g. 100:      100
Largest band size in the range? e.g. 1000:      1000
Minimum RFI cutoff value, e.g. 0.09%?           0.09

Window size (bp)? (0.5, 1, 1.5, 2)...           1
Shift size (bp)? (0.1, 0.5, 1).......           0.1

Plot the results? (y/n)                         y
Outputting to text files? (y/n)                 y
Checking for existence of bands in the defined peak range: 100-1000 ...
OK. No problem of sizes
------------------------------------------------------------
please wait...

Best bin frame is:  0.2  (highest mean correlations)
Max OTU number for frame:  0.5 0.6      (28 OTUs)
(Chosen parameters: WS=1, Shift=0.1)
The results are available in the following object: Result

Result summary for each frame:
                0    0.1   0.2   0.3  0.4   0.5   0.6   0.7   0.8   0.9
Correlations  0.76  0.85  0.85  0.85  0.8  0.56  0.54  0.59  0.67  0.67
NberOTUs     25.00 22.00 23.00 23.00 24.0 28.00 28.00 26.00 24.00 23.00
> |
```

This time, the script did not stop because no size problems were encountered and thus the calculations were done.

(Note that the script will also send an error message and stop if the RFI cutoff value is set too high. In the latter case, it would remove too many peaks for calculations to be correctly performed).

5. Analyzing the results

The console above indicates the best bin frame (starting at position 0.2 bp) out of the 10 frames being compared (the shift value was 0.1 bp for a window size of 1 bp) based on the highest mean correlations among samples (see reference below for more information). The maximum number of OTU is also reported for all the matching frames (e.g. for frames starting at 0.5 and 0.6), as well as their corresponding number of Operational Taxonomic Units (OTUs).



Those results are also plotted if you indicated “y” for the plotting option. You can export the graphics to your favorite picture editing software by right-mouse clicking and saving.

Back to the console, all frames and calculations are available in the current workspace and are aggregated in the `Result` object that you created by typing D2res. By typing `D2res` (but this is generally not recommended because of the size of the ouput), you will see all the information stored for your analysis.

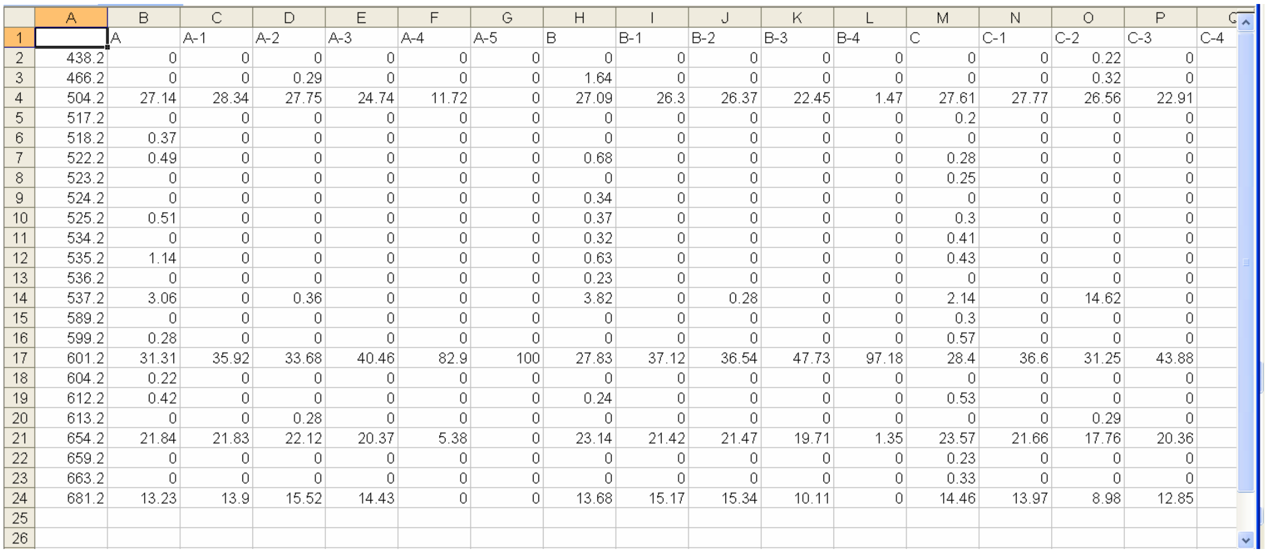
If you rather want to check a particular table, it is more convenient to type one of the following commands:

Commands	***Definition***
`D2res[[1]]`	D table with the calculated RFI values in the last column
`D2res[[2]]`	Unbinned table of samples by peak sizes
`D2res[[3]]`	List of all binned frames
`D2res[[4]]`	List of all correlation values among samples for each binning frame
`D2res[[5]]`	Mean correlation per binning frame
`D2res[[6]]`	Best bin frame identified
`D2res[[7]]`	Summary of sample correlation and OTU number for each frame

Note that the Relative Fluorescence Intensity (RFI) values are calculated for the defined size range, and thus will vary accordingly. The most important results were also saved in the current working directory if you indicated "`y`" for the outputting option.

The files are:	***which correspond to:***
output_D_RFI.txt	Result[[1]]
output_bandlistRFI.txt	Result[[2]
output_bandlist01.txt	Result[[2]] converted to presence-absence instead of RFI
output_best.binned.table0.2.txt	The best binned frame identified. Note that the ending of the file name will change (i.e. "0.2") depending on the best solution for the data.
output_summary.txt	Result[[7]]

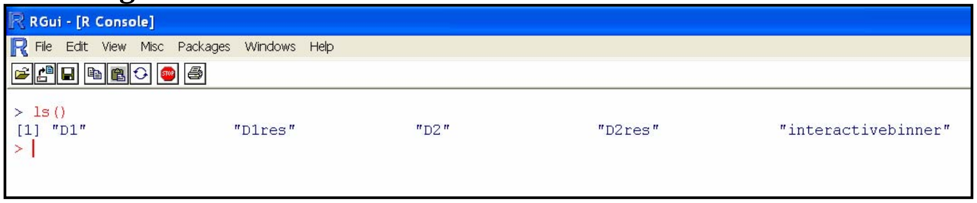
You can export the data to Excel and convert the text to data (Menu bar: Data/Text To Columns/Delimited/Space box checked). Note that the OTUs are indicated only by the start of the bin.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1		A	A-1	A-2	A-3	A-4	A-5	B	B-1	B-2	B-3	B-4	C	C-1	C-2	C-3	C-4
2	438.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0.22	0	
3	466.2	0	0	0.29	0	0	0	1.64	0	0	0	0	0	0	0.32	0	
4	504.2	27.14	28.34	27.75	24.74	11.72	0	27.09	26.3	26.37	22.45	1.47	27.61	27.77	26.56	22.91	
5	517.2	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0	
6	518.2	0.37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	522.2	0.49	0	0	0	0	0	0.68	0	0	0	0	0.28	0	0	0	
8	523.2	0	0	0	0	0	0	0	0	0	0	0	0.25	0	0	0	
9	524.2	0	0	0	0	0	0	0.34	0	0	0	0	0	0	0	0	
10	525.2	0.51	0	0	0	0	0	0.37	0	0	0	0	0.3	0	0	0	
11	534.2	0	0	0	0	0	0	0.32	0	0	0	0	0.41	0	0	0	
12	535.2	1.14	0	0	0	0	0	0.63	0	0	0	0	0.43	0	0	0	
13	536.2	0	0	0	0	0	0	0.23	0	0	0	0	0	0	0	0	
14	537.2	3.06	0	0.36	0	0	0	3.82	0	0.28	0	0	2.14	0	14.62	0	
15	589.2	0	0	0	0	0	0	0	0	0	0	0	0.3	0	0	0	
16	599.2	0.28	0	0	0	0	0	0	0	0	0	0	0.57	0	0	0	
17	601.2	31.31	35.92	33.68	40.46	82.9	100	27.83	37.12	36.54	47.73	97.18	28.4	36.6	31.25	43.88	
18	604.2	0.22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	612.2	0.42	0	0	0	0	0	0.24	0	0	0	0	0.53	0	0	0	
20	613.2	0	0	0.28	0	0	0	0	0	0	0	0	0	0	0.29	0	
21	654.2	21.84	21.83	22.12	20.37	5.38	0	23.14	21.42	21.47	19.71	1.35	23.57	21.66	17.76	20.36	
22	659.2	0	0	0	0	0	0	0	0	0	0	0	0.23	0	0	0	
23	663.2	0	0	0	0	0	0	0	0	0	0	0	0.33	0	0	0	
24	681.2	13.23	13.9	15.52	14.43	0	0	13.68	15.17	15.34	10.11	0	14.46	13.97	8.98	12.85	
25																	
26																	

For instance, OTU “438.2” corresponds to peak size in [438.2-439.2[because of the window size of 1 bp. Bins without any peak present were removed from the table, so the actual first OTU may not start at 100.2 (and this is the case here).

6. Exiting from R



```
> ls()
[1] "D1"                    "D1res"                 "D2"                    "D2res"                 "interactivebinner"
> |
```

The data are stored in the current workspace and you can save them via the File\save workspace option in the menu bar. Note that typing ls() lists all objects currently available in your session. You can choose to save the objects or not for future work before closing the R console.

How to cite the script?

Ramette, A. (2008) Quantitative molecular community fingerprinting for estimating the abundance of operational taxonomic units in natural microbial communities. *submitted*