msap (v. 0.1.1) - User's Guide

Andres Perez-Figueroa

June 15, 2012

1 Introduction

msap provides a deep analysis of epigenetic variation starting from a binary data matrix indicating the presence or absence of EcoRI-HpaII and EcoRI-MspI fragments, typical of MSAP technique. After compare the data from both ezyme combinations, the program determines if each fragment is susceptible of methylation (representative of epigenetic variation) or if there is no evidence of methylation (representative of genetic variation). Different analyses of the variation (genetic and epigenetic) among user-defined groups of samples are then performed, as well as the classification of the methylation ocurrences in those groups. Statistical testing provide support to the analyses. A comprehensive report of the analyses and several useful plots could help researchers to asses the epigenetic variation in their experiments using MSAP.

The package is intended to be easy to use even for those people non-familiar to the R environment. Advanced users could take advantange of available source code to adapt msap for more complex analyses.

2 Installing *msap*

You can install *msap* automatically from a R session. To install the last stable version from CRAN (Not available yet):

```
> install.packages("msap")
```

To get the last daily development version from R-Forge:

```
> install.packages("msap", repos="http://R-Forge.R-project.org")
```

The above instructions should install *msap* and all required dependecies.

3 Preparation of data

In order to use *msap* to analyse your results from a MSAP experiment, you need to provide a data file with a binary matrix (1/0) indicating the presence

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	I
1				m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m1
2	Pop1	a1	HPA	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
3	Pop1	a1	MSP	0	0	0	1	1	0	0	1	1	1	0	0	0	0	0	0	1	
4	Pop1	a2	HPA	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
5	Pop1	a2	MSP	0	0	1	0	0	1	1	1	1	1	1	0	0	1	0	1	0	
6	Pop1	a3	HPA	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
7	Pop1	a4	MSP	1	1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	1	
8	Pop1	a5	HPA	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
9	Pop1	a5	MSP	0	0	1	0	0	0	1	0	1	1	1	0	0	1	0	1	0	
10	Pop1	a6	HPA	1	1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	1	
11	Pop1	a6	MSP	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0	0	0	
12	Pop1	a7	HPA	0	1	0	1	0	0	0	1	1	1	1	0	0	1	0	0	0	
13	Pop1	a7	MSP	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	
14	Pop2	b1	HPA	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
15	Pop2	b1	MSP	0	0	0	0	1	0	0	0	1	1	0	1	0	0	0	0	0	
16	Pop2	b2	HPA	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
17	Pop2	b2	MSP	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
18	Pop2	b3	HPA	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	
19	Pop2	b3	MSP	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	
20	Pop2	b4	HPA	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
21	Pop2	b4	MSP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22	Pop2	b5	HPA	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	
23	Pop2	b5	MSP	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	
24	Pop2	b6	HPA	0	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	
25	Pop2	b6	MSP	0	1	0	0	1	1	0	1	1	0	1	0	0	1	0	0	0	
26	Pop3	c1	HPA	0	1	0	1	1	0	0	1	1	1	0	0	0	1	0	0	1	
27	Pop3	c1	MSP	0	0	1	1	0	0	0	1	1	1	1	0	0	0	0	0	0	
28	Pop3	c2	HPA	0	1	0	1	1	0	0	1	1	1	0	0	0	0	0	1	0	
29	Pop3	c2	MSP	0	0	0	1	1	0	0	1	1	1	1	0	0	1	0	0	1	
30	Pop3	c3	HPA	1	1	0	1	1	0	0	1	1	1	1	0	0	0	1	0	1	
31	Pop3	c3	MSP	1	1	0	1	0	0	1	1	1	1	0	0	0	0	1	0	1	

Figure 1: Data format as seen in a spreadsheet for edition

or absence of EcoRI-HpaII and EcoRI-MspI fragments in a bunch of samples of two or more populations/groups. Data file should be a .csv file with markers as columns and two rows by sample, one for each isoschizomer reaction. File could be edited in the a spreadsheet of your choice (see Figure 1) and then saved as csv (with ',' as field separator). The final text file should look like Figure 2 if opened in a text editor. The first row should include the markers name/references. The first column should provide the label for the group where the sample is included, with the aim to make comparisons between different gruops. Second column is reserved for an arbitrary label (i.e. to name the sample). Third column should identify the isoschizomer with 'HPA' or 'MSP'.

4 Executing *msap*

We start by loading the msap package into an R session.

```
> library(msap)
```

It is highly recommended to change the working directory to that where datafile is located. Windows users can use the menu item 'File>Change dir' and choose the appropiate folder. To change the working directory within an R console run the command `setwd(dir)` where *dir* is the absolute path to the directory. The output files created by *msap* will be save in that working directory.

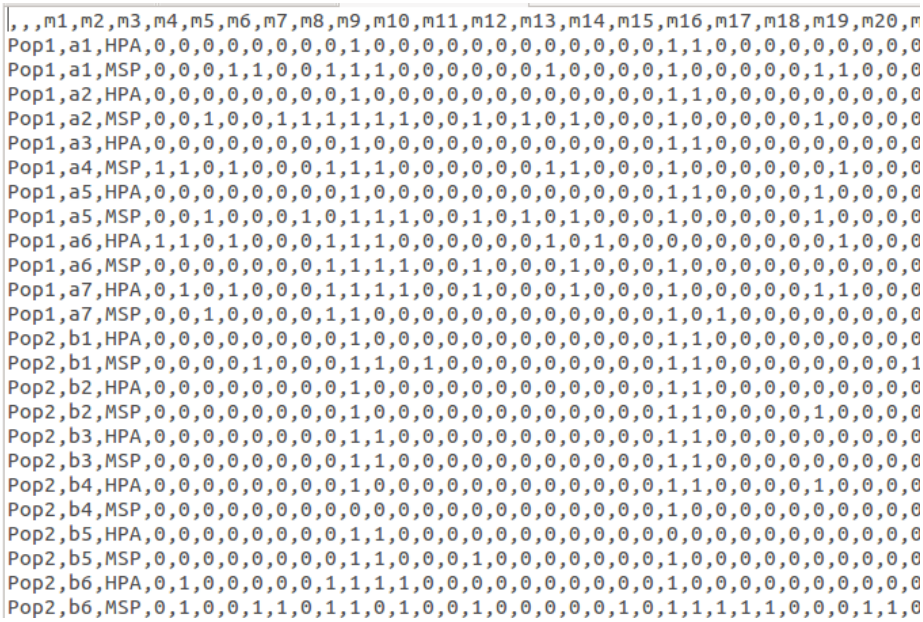

Figure 2: Final data format in the .csv file

Once we are in the rigth working directory with an appropiate data file, we can run all analyses of *msap* with a single command:

```
> msap("example.csv",name="Example")

msap - Statistical analysis for Methilation-Sensitive Amplification Polimorphism data

Reading  example.csv
Number of Methylation-Susceptible Loci (MSL):  337
Number of No Methylated Loci (NML):  81

Report of methylation levels
                                                      Pop1
HPA+/MSP+ (Unmethylated)                            0.1627
HPA+/MSP- (Hemimethylated)                          0.1548
HPA-/MSP+ (Internal cytosine methylation)           0.2171
HPA-/MSP- (Full methylation or absence of target)   0.4654
                                                      Pop2
HPA+/MSP+ (Unmethylated)                            0.1573
HPA+/MSP- (Hemimethylated)                          0.1385
HPA-/MSP+ (Internal cytosine methylation)           0.1855
HPA-/MSP- (Full methylation or absence of target)   0.5188
                                                      Pop3
HPA+/MSP+ (Unmethylated)                            0.1573
```

```
HPA+/MSP- (Hemimethylated)                        0.1526
HPA-/MSP+ (Internal cytosine methylation)         0.1509
HPA-/MSP- (Full methylation or absence of target) 0.5392

Shannon's Diversity Index
MSL: I =  0.5491545   (SD:  0.1270955 )
NML: I =  0.2122527   (SD:  0.02776546 )
Wilcoxon rank sum test with continuity correction :  W  =  8389  ( P < 0.0001 )

Analysis of MSL

Performing AMOVA
AMOVA TABLE         d.f.         SSD                MSD                Variance
among groups         2          0.4237          0.2118          0.01647
within groups         16          1.725          0.1078          0.1078
Total                 18          2.149          0.1194

Phi_ST =  0.1325   (P= 0.0015 )

Pairwise Phi_ST
-----------------------------------
Pop1  -  Pop2 :  0.08651           (P= 0.0399 )
Pop1  -  Pop3 :  0.2586           (P= 0.0013 )
Pop2  -  Pop3 :  0.0285           (P= 0.2088 )

Analysis of NML

Performing AMOVA
AMOVA TABLE         d.f.         SSD                MSD                Variance
among groups         2          0.01177          0.005883          -0.0002831
within groups         16          0.1227          0.007671          0.007671
Total                 18          0.1345          0.007472

Phi_ST =  -0.03832   (P= 0.9751 )

Pairwise Phi_ST
-----------------------------------
Pop1  -  Pop2 :  -0.06015           (P= 0.9793 )
Pop1  -  Pop3 :  0.02681           (P= 0.1466 )
Pop2  -  Pop3 :  -0.06821           (P= 0.9868 )
```

Aditionally to the on-screen report, the following figures are produced and stored in .png files:

- A boxplot with the distribution of Shannon's diversity indices in both MSL and NML (Figure 3)

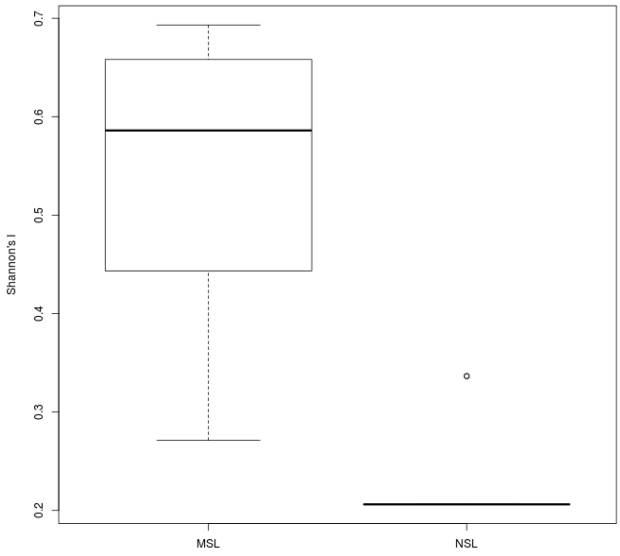

Figure 3: Boxplot comparing Shannon's Diversity Index in MSL and NML

- A plot with the representation of Principal Coordinate Analysis (PCoA) for epigenetic (MSL) differentiation between groups. (Figure 4)
- A plot with the representation of Principal Coordinate Analysis (PCoA) for genetic (NML) differentiation between groups. (Figure 5)

4.1 Further options

In the previous section, the basic use of msap was described. However, it is possible to set some different options in the program if passed as arguments to the msap() function.

Here is the full usage of msap() function including all the arguments and their default values (if applicable):

```
msap(datafile, name=datafile, uninformative=TRUE, nDec=4)
```

datafile String containing the url of the csv file with the data. Required.

name a name for the dataset to be included in the output files. By default, the name of the given datafile is used.

uninformative A logical value determining how to deal with HPA-/MSP- pattern. 'FALSE' assumes that HPA-/MSP- (no band for both isoschizomers)

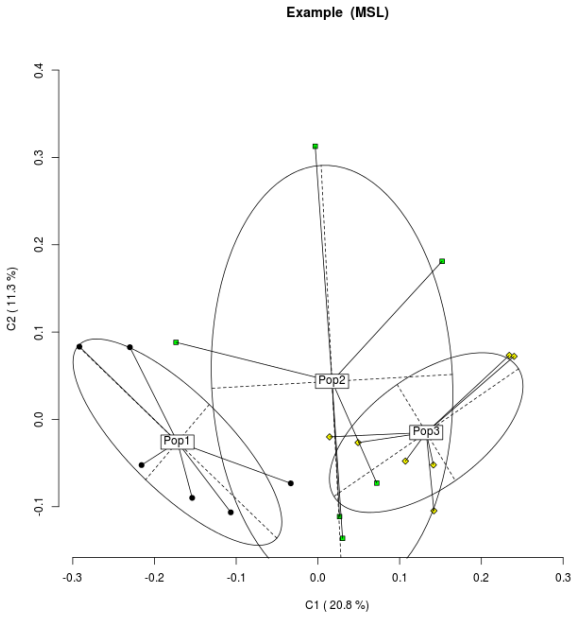

Figure 4: Representation of Principal Coordinate Analysis (PCoA) for epigenetic (MSL) differentiation between groups. The first two coordinates (C1 and C2) are shown with the percentage of variance explained by them. Diferente point types represent individuals from different groups. Group labels show the centroid for the points cloud in each group. Ellipses represent the average dispersion of those poins aroun their centre. The long axis of the ellipse shows the direction of maximum dispersion and the short axis, the direction of minimum dispersion

pattern represents full methylation of cytosines in the target, while 'TRUE' (default value) consider that pattern as uninformative as could be caused by a missing target (mutation). See 'Details' below

nDec number of digits of precision for floating point output.

5 Session Info

This document was created using the following:

```
> sessionInfo()

R version 2.15.0 (2012-03-30)
Platform: i686-pc-linux-gnu (32-bit)
```

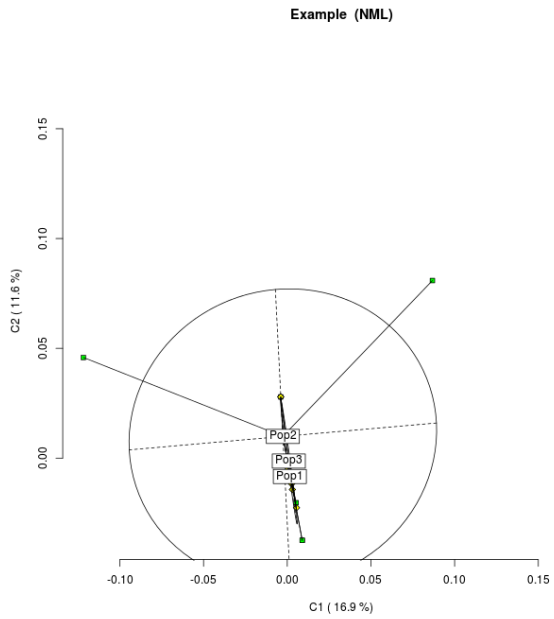

Figure 5: Representation of Principal Coordinate Analysis (PCoA) for genetic (NML) differentiation between groups.

```
locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] grid      stats     graphics  grDevices utils
[6] datasets  methods   base

other attached packages:
[1] cba_0.2-9      proxy_0.4-7    pegas_0.4-2
[4] adegenet_1.3-4 MASS_7.3-16    ape_3.0-3
[7] scrime_1.2.8   ade4_1.4-17    msap_0.1.1

loaded via a namespace (and not attached):
[1] gee_4.13-18    lattice_0.20-6 Matrix_1.0-6
[4] nlme_3.1-103   tools_2.15.0
```

References