msap (v. 0.1-2) - User's Guide

Andres Perez-Figueroa

June 19, 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 enzyme 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 and differentiation (genetic and epigenetic) among user-defined groups of samples are then performed, as well as the classification of the methylation occurrences in those groups. A comprehensive report of the analyses and several useful plots could help researchers to asses the epigenetic variation in their experiments using MSAP. Standard AFLP data is also suitable to be analyzed. All analyses follow a band-based strategy (Bonin *et al.*, 2007)

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

2 R basics. All you need to know about R to run *msap*

The only knowledge required for installing and running *msap* is about how to open an R session in your computer. R is a statistical programming language that provides many built in functions for performing statistical analysis and is also fexible to allow users to write their own functions.

R can be downloaded and installed for free from the website http://cran.r-project.org where detailed instructions for installing R on any operating system are provided. Accessing R is different for every operating system. For windows users, simply double click the R icon that is created after installation. For Mac users, you can double click the R icon under your Applications menu. On Linux, in the terminal window simply type R at the command prompt and R will be opened within the terminal window.

When you open R, no matter the operating system you are using, you will see the command prompt symbol `>` which simply means that R is waiting for you to give it a command. To quit R, simply type `q()` in the command prompt and R will ask you if you want to save the workspace before quitting. And that's all.

3 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.

4 Preparation of data

In order to use *msap* for analyzing your results from a MSAP experiment, you need to provide a data file with a binary matrix (1/0) indicating the presence 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 groups. Second column is reserved for an arbitrary label (i.e. to name the sample). Third column should identify the isoschizomer with 'HPA' or 'MSP'. If you want to analyze a standard AFLP dataset the datafile format is the same, but the program will ignore content of third column and treat all rows as independent samples.

5 Running *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

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1				m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17
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

```
,,,m1,m2,m3,m4,m5,m6,m7,m8,m9,m10,m11,m12,m13,m14,m15,m16,m17,m18,m19,m20,m
Pop1,a1,HPA,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0
Pop1,a1,MSP,0,0,0,1,1,0,0,1,1,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,1,1,0,0,0
Pop1,a2,HPA,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0
Pop1,a2,MSP,0,0,1,0,0,1,1,1,1,1,1,0,0,1,0,1,0,1,0,0,0,1,0,0,0,0,0,1,0,0,0,0
Pop1,a3,HPA,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0
Pop1,a4,MSP,1,1,0,1,0,0,0,1,1,1,0,0,0,0,0,0,1,1,0,0,0,1,0,0,0,0,0,0,1,0,0,0
Pop1,a5,HPA,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,1,0,0,0,0,0
Pop1,a5,MSP,0,0,1,0,0,0,1,0,1,1,1,0,0,1,0,1,0,1,0,0,0,1,0,0,0,0,0,1,0,0,0,0
Pop1,a6,HPA,1,1,0,1,0,0,0,1,1,1,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0
Pop1,a6,MSP,0,0,0,0,0,0,0,0,1,1,1,1,0,0,1,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0
Pop1,a7,HPA,0,1,0,1,0,0,0,1,1,1,1,1,0,0,1,0,0,0,1,0,0,0,1,0,0,0,0,0,1,1,0,0,0
Pop1,a7,MSP,0,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0
Pop2,b1,HPA,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0
Pop2,b1,MSP,0,0,0,0,1,0,0,0,1,1,0,1,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,1
Pop2,b2,HPA,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0
Pop2,b2,MSP,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,1,0,0,0,0,0
Pop2,b3,HPA,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0
Pop2,b3,MSP,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0
Pop2,b4,HPA,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,1,0,0,0,0,0
Pop2,b4,MSP,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
Pop2,b5,HPA,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Pop2,b5,MSP,0,0,0,0,0,0,0,0,1,1,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
Pop2,b6,HPA,0,1,0,0,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
Pop2,b6,MSP,0,1,0,0,1,1,0,1,1,0,1,0,0,1,0,0,0,0,0,1,0,1,1,1,1,1,0,0,0,1,1,0
```

Figure 2: Final data format in the .csv file

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.

Once we are in the rigth working directory with an appropiate data file, we can run all analyses of *msap* with a single command (change "example.csv" by the name of your datafile, keeping the quotes, and change "Example" by custom name to identify your data):

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

msap - Statistical analysis for Methilation-Sensitive Amplification Polimorphism data

Reading  example.csv
Number of loci:  701
Number of samples/individuals:  38
Number of groups/populations:  4
Number of Methylation-Susceptible Loci (MSL):  494
Number of No Methylated Loci (NML):  207

Number of polymorphic MSL:  257  ( 52 % of total MSL)
Number of polymorphic NML:  49  ( 24 % of total NML)


Shannon's Diversity Index
MSL: I =  0.5836821   (SD:  0.1035363 )
NML: I =  0.3299215   (SD:  0.1078034 )
Wilcoxon rank sum test with continuity correction :  W  =  11710  ( P < 0.0001 )


****************************
Analysis of MSL
Report of methylation levels
                                                    pop1
HPA+/MSP+ (Unmethylated)                          0.1917
HPA+/MSP- (Hemimethylated)                        0.1706
HPA-/MSP+ (Internal cytosine methylation)         0.2156
HPA-/MSP- (Full methylation or absence of target) 0.4221
                                                    pop2
HPA+/MSP+ (Unmethylated)                          0.1688
HPA+/MSP- (Hemimethylated)                        0.1432
HPA-/MSP+ (Internal cytosine methylation)         0.1430
HPA-/MSP- (Full methylation or absence of target) 0.5450
                                                    pop3
HPA+/MSP+ (Unmethylated)                          0.1601
HPA+/MSP- (Hemimethylated)                        0.1599
HPA-/MSP+ (Internal cytosine methylation)         0.2298
```

```
HPA-/MSP- (Full methylation or absence of target) 0.4502
                                                    pop4
HPA+/MSP+ (Unmethylated)                          0.1670
HPA+/MSP- (Hemimethylated)                        0.1281
HPA-/MSP+ (Internal cytosine methylation)         0.1589
HPA-/MSP- (Full methylation or absence of target) 0.5460

Performing AMOVA
AMOVA TABLE         d.f.         SSD                MSD                Variance
among groups         3          1.332          0.4439          0.03146
within groups         34          4.961          0.1459          0.1459
Total                 37          6.293          0.1701

Phi_ST =  0.1773   (P<0.0001)

Pairwise Phi_ST
-------------------------------------
pop1  -  pop2 :  0.2971          (P<0.0001)
pop1  -  pop3 :  0.05566          (P= 0.0218 )
pop1  -  pop4 :  0.2341          (P<0.0001)
pop2  -  pop3 :  0.203          (P= 2e-04 )
pop2  -  pop4 :  0.09294          (P= 0.0017 )
pop3  -  pop4 :  0.1518          (P<0.0001)


******************************
Analysis of NML

Performing AMOVA
AMOVA TABLE         d.f.         SSD                MSD                Variance
among groups         3          0.3587          0.1196          0.006628
within groups         34          1.931          0.05679          0.05679
Total                 37          2.29          0.06188

Phi_ST =  0.1045   (P<0.0001)

Pairwise Phi_ST
-------------------------------------
pop1  -  pop2 :  0.2213          (P<0.0001)
pop1  -  pop3 :  0.02315          (P= 0.1173 )
pop1  -  pop4 :  0.1757          (P<0.0001)
pop2  -  pop3 :  0.08653          (P= 0.0021 )
pop2  -  pop4 :  0.0206          (P= 0.07299 )
pop3  -  pop4 :  0.07359          (P= 0.009299 )
```

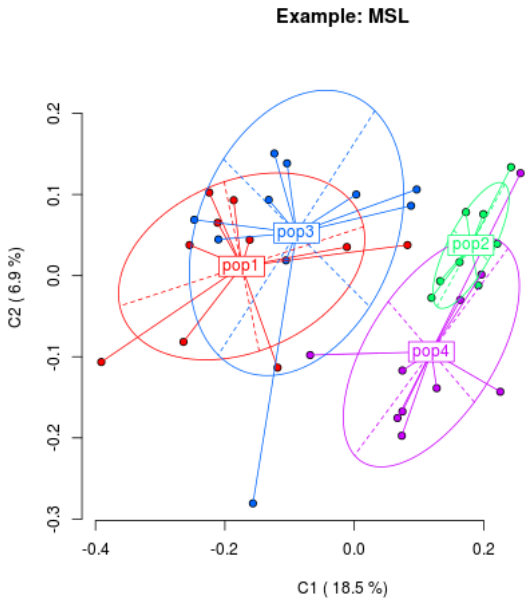

Figure 3: Representation of Principal Coordinate Analysis (PCoA) for epigenetic (MSL) differentiation between groups. The first two coordinates (C1 and C2) are shown with the percentage of variance explained by them. Diferente point types represent individuals from different groups. Group labels show 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.

In addition to the above text report, *msap* produces some exploratory figures that are directly saved into .png files:

- A plot of Principal Coordinate Analysis (PCoA), see Figure 3, showing the first two axis. They are saved as '<*name*>-MSL.png' and '<*name*>-NML.png' for MSL and NML respectively, where <*name*> represents the name passed as argument in the calling to msap function.
- A neighbor-joining tree of all samples (see Figure 4) saved as '<*name*>-MSL-NJ.png' and '<*name*>-NML-NJ.png' for MSL and NML respectively

5.1 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.

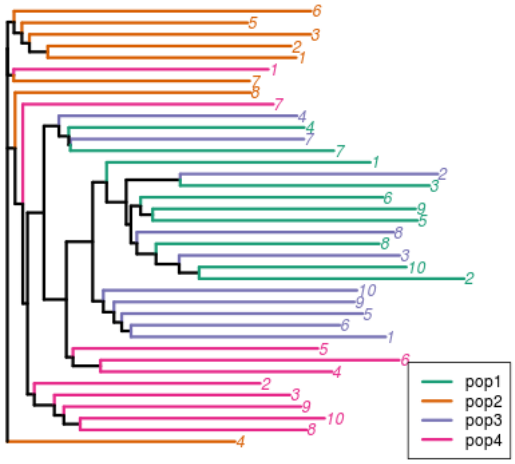

Figure 4: Neighbor-Joining tree of all samples (numbered labels at the tips) for epigenetic (MSL) distances. Colors represent different groups/populations.

Here is the full usage of msap() function including all the arguments and their default values (if applicable). Except for the 'datafile' that is required, all other arguments are optional.

```
msap(datafile, name=datafile, uninformative=TRUE,
nDec=4, meth=TRUE, rm.redundant=TRUE,
rm.monomorphic=TRUE, do.pcoa=TRUE, do.shannon=TRUE,
do.amova=TRUE, do.pairwisePhiST=TRUE, do.cluster=TRUE,
use.groups=NULL)
```

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) 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.

meth Logical value switching between MSAP ('TRUE') and standard AFLP ('FALSE') analysis. The difference lies in that for AFLP (meth=FALSE) the 'enzyme' column is ignored and every row in data represent an independent sample, without combination of data.

rm.redundant Not implemented yet.

rm.monomorphic Logical value switching between the removal ('TRUE', by default) of monomorphic fragments (defined as those with only one state or just one ocurrence of the second state across the whole dataset) after data transformation.

do.pcoa Option switcher for doing a Principal Coordinate Analysis for variation between groups. TRUE by default.

do.shannon Option switcher for Shannon's Diversity Index comparison between MSL and NML.

do.amova Option switcher for doing an AMOVA for differentiation between groups. TRUE by default.

do.pairwisePhiST Logical value switching between the calculation of the pairwise Φ_{st} between pairs of groups/populations ('TRUE' by default) or skip it ('FALSE').

do.cluster Calculates and plots a Neighbour-Joining tree ('TRUE' by default) or skip it ('FALSE').

use.groups Gives the groups/populations/treatments of the datafile to be analysed. By default all groups are considered into de the analysis. To provide a subset of the groups a vector should be passed with the names of groups to be included. For example, in a datafile with 5 groups (Control, pop1, pop2, pop3 and pop4) we are interested only in Control and pops 1 and 3. Then, msap should be called with 'use.groups=c('Control','pop1','pop3')'.

5.2 What if I need another kind of analysis for my MSAP data?

The analyses currently provided by *msap* are limited but that does not mean that further analysis could be added in the future. I try to keep the package updated to allow exploratory assays of epigenetic diversity and differentiation, focused on the field of evolutionary ecology. If you are trying to analyse your MSAP data and options in *msap* do not fit your requeriments, then you have three alternatives to get your tasks done by *msap*:

- If you have programming skills or experience with R, then you can get the source code of *msap* and adapt it to your needs. That is the main advantage of open source software!
- If you have programming skills or experience with R, and want to collaborate with me to expand *msap*, then do not hesitate to contact me for joing to the development team.
- If you are not familiar with R or do not feel confident to make code yourself, then use the support tracker to request new features in *msap*.

6 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)

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

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
```