MSAP_calc.r - R-functions for transformation and analysis of MSAP data

Walter Durka

MSAP_calc.r v. 1.3, 07.02.2017
Updated version available online at http://www.ufz.de/index.php?en=816
walter.durka@ufz.de

Introduction

Methylation sensitive amplification polymorphism (MSAP) is a method to identify methylation polymorphisms. It is similar to amplified fragment length polymorphism (AFLP), i.e. two restriction enzymes are used to produce DNA fragments which are PCR-amplified with selective primers. MSAP uses *Eco*RI as rare cutter and *Hpa*II and *Msp*I as frequent cutter. *Hpa*II and *Msp*I differ in their susceptibility to DNA methylation. Essentially, two analyses are performed for each sample, one with *Eco*RI + *Hpa*II and one with *Eco*RI + *Msp*I. After fragment analysis, fragments are scored as present ("1") or absent ("0"). Thus, depending on presence of a fragment in both analyses, a combined score for the *Hpa*II / *Msp*I analyses is obtained: 1/1, 1/0, 0/1 and 0/0.

	condition I	condition II	condition III	condition IV
*Hpa*II/*Msp*I - fragments	1/1	0/1	1/0	0/0
methylation state of fragment	unmethylated	^{HMe}CG & ^{Me}CG	^{HMe}CCG	full methylation of external or both cytosines or fragment absence
primary coding (Paun et al. 2010)	2	-1	1	0

In order to analyse such data, these MSAP epigenotypes are usually transformed into a data matrix of dominant (1/0) markers that is amenable for traditional population genetic analysis. Various transformation schemes have been developed (Herrera & Bazaga 2010; Lira-Medeiros *et al.* 2010; Paun *et al.* 2010; Salmon *et al.* 2008; Vergeer *et al.* 2012) and others are possible, all of which are discussed in detail in the accompanying paper (Schulz *et al.* 2013). One of the newly proposed scoring schemes ("Mixed 2") is used (Schulz *et al.* 2014).

MSAP_calc offers functions to transform primary MSAP data into binary epigenetic loci and to calculate descriptive parameters of epigenetic variation using the R environment (R Core Team 2012).

Input data format

MSAP_calc expects data formatted as tab-delimited ASCII txt file. The first line must contain the names for the variables and the data are given in lines 2 to N_sample+1. In each line, the three first columns contain population-ID, sample-ID, restriction enzyme (“H” or “M”) and columns 4 to N_markers+3 contain presence (“1”) or absence (“0”) of fragments.
For each sample, two lines of data are expected, with identical population-ID and sample-ID, one with “H” and one with “M”-data (if this is not the case, the function will not work properly). Missing values are not allowed and characters other than ”0” and “1” will lead to false results. An example data set is given in “MSAP_data.txt”.

MSAP_calc input file for two individuals from two populations and two markers

populationID	sampleID	restriction_enzyme	AAC_CA_135	AAC_CA_163
1	1	H	1	0
1	1	M	0	0
1	2	H	1	1
1	2	M	1	0
2	1	H	1	1
2	1	M	1	1
2	2	H	0	0
2	2	M	0	1

Functions

Extract_MSAP_epigenotypes

```
Extract_MSAP_epigenotypes <- function( inputfile="MSAP_data.txt",
                                       Epicode = "Mix1",
                                       outputfile="MSAP_out.txt",
                                       MinPoly=1,
                                       delete.monomorphic.loci=TRUE)
```

This function reads the data from `inputfile`, and transforms the data into a 0/1 matrix according to various transformation schemes (`Epicode`, see Table 1). Transformed data are delivered and written to a file `outputfile`.

From the primary data matrix, loci are deleted if they have less then `MinPoly` polymorphisms in the whole data set. From the final matrix, monomorphic loci are deleted if `delete.monomorphic.loci=TRUE`.

All loci are renamed by adding a one-character prescript ("e") to the original locusID to indicate that data is different from the original HpaII/MspI scores. For those transformations ("Paun", "Mix1", "Mix2") that extract multiple epiloci from one original fragment, the locus prescript indicates the type of epilocus: "u" for unmethylated, "m" for ^{HMe}CG & ^{Me}CG, "h" for ^{HMe}CCG, "M" for methylation (either "m" or "h").

Note that this function runs slow and may take a few minutes for large datasets.

Table 1. MSAP data transformation schemes

Epicode	**Types of loci**	**condition I**	**II**	**III**	**IV**	**Locus**	**Ref**
	H/M	1/1	0/1	1/0	0/0	**prescript**	
"Salmon"	epigenetic	0	1	1	0	e	1
"Vergeer"	epigenetic	0	1	1	NA	e	2
"Herrera"	meth.sensitive (epigenetic)	0	1	1	NA	e	3
"Lira-M1"	epigenetic	1	0	loci excluded	0	e	4
"Lira-M2"	epigenetic	1	0	0	0	e	5
"Paun"	unmethylated	1	0	0	0	u	6
	^{HMe}CG & ^{Me}CG	1	1	0	0	m	
	^{HMe}CCG	1	0	1	0	h	
"Mix1"	unmethylated	1	0	0	0	u	5
	methylated total	0	1	1	0	M	
"Mix2"	unmethylated	1	0	0	0	u	5
	^{HMe}CG & ^{Me}CG	0	1	0	0	m	
	^{HMe}CCG	0	0	1	0	h	

Ref.: [1]Salmon et al. 2008; [2]Vergeer et al. 2012; [3]Herrera & Bazaga 2010 ; [4]Lira-Medeiros et al. 2010; [5]Schulz et al. 2013; [6]Paun et al. 2010

descriptive_parameters

```
descriptive_parameters <- function ( inputfile="MSAP_Mix2.txt",
                                     outputfile="MSAP_Mix2_descr.txt",
                                     AppendOutput=FALSE)
```

This function reads a file `inputfile` produced by `Extract_MSAP_epigenotypes` and calculates descriptive parameters at the population level, which are delivered and written to `outputfile` as tab-delimited ASCSII file (which is appended or overwritten, depending on `AppendOutput`). The following parameters are calculated:

0. PopID	population identifier
1. N_samples	number of samples per population
2. N_markers_total	total number of markers in data set
3. N_markers_pop	number of markers present (with at least one "1"-score) per population
4. N_markers_poly	number of markers polymorphic per population
5. Pc_markers_poly	percentage markers polymorphic per population
6. Mean_N_1scores	mean number of "1"-scores per population
7. N_private_markers	number of private markers per population, i.e. markers that only occur in this population
8. Shannon_diversity	Mean across loci of the Shannon index of phenotypic diversity $H'_{epi} = -\sum p_i \log_2 p_i$

For all epigenetic transformations that distinguish between different types of epigenetic loci (see Tab. 1), parameters 2 to 8 are additionally analyzed for each of these groups separately. The parameter names are amended with a prescript indicating the type of epiloci: "u" for unmethylated, "m" for ^{HMe}CG & ^{Me}CG , "h" for ^{HMe}CCG, "M" for methylation-variable loci.

Table 2. MSAP_calc output file for Mixed Scoring 2 with four different sets of descriptive parameters for all, ^{HMe}CCG, ^{HMe}CG & ^{Me}CG and unmethylated loci.

Descriptive parameters of populations calculated from MSAP_Mix2.txt

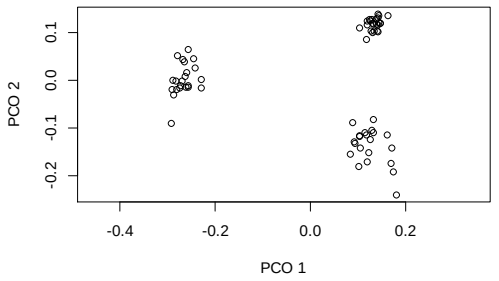
PopID	N_samples	N_markers_total	N_markers_pop	N_markers_poly	Pc_markers_poly	Mean_N_1scores	N_private_markers	Shannon_diversity	h_N_markers_total	h_N_markers_pop	h_N_markers_poly	h_Pc_markers_poly	h_Mean_N_1scores	h_N_private_markers	h_Shannon_diversity	m_N_markers_total	m_N_markers_pop	m_N_markers_poly	m_Pc_markers_poly	m_Mean_N_1scores	m_N_private_markers	m_Shannon_diversity	u_N_markers_total	u_N_markers_pop	u_N_markers_poly	u_Pc_markers_poly	u_Mean_N_1scores	u_N_private_markers	u_Shannon_diversity
R7	22	286	184	143	50	7.8	44	0.27	53	32	30	57	5	11	0.32	105	71	58	55	9.1	16	0.28	128	81	55	43	8	17	0.24
R8	24	286	172	110	38	9.2	15	0.19	53	26	25	47	2.9	5	0.25	105	63	43	41	9.4	4	0.19	128	83	42	33	12	6	0.16
R9	21	286	212	181	63	7.9	44	0.34	53	35	34	64	2.9	10	0.31	105	78	73	70	9.1	18	0.38	128	99	74	58	8.8	16	0.32

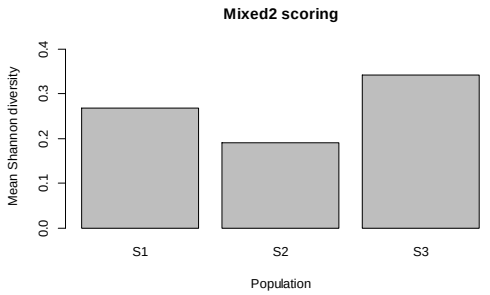
All loci | HMeCCG | HMeCG & MeCG | unmethylated

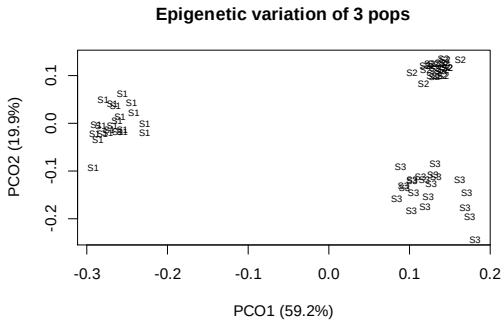
A session with MSAP_calc.r

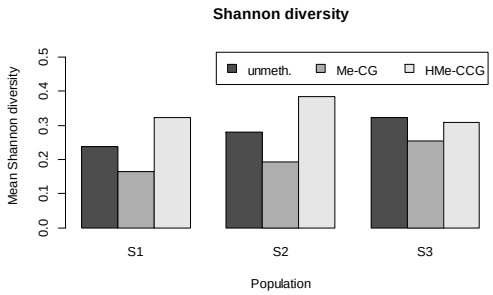
```
#############################################################################
rm(list=ls())
setwd('E:/A/MSAP')       # go to working directory
source("MSAP_calc_1_2.r")    # load functions from MSAP_calc_1_2.r

# Read primary MSAP data from MSAP_data.txt and transform according to "Mixed
# Scoring 2" scheme of Schulz et al.
# Transformed data are available as data.frame "d" and output to "MSAP_Mix2.txt"
# --> be patient ! Takes 1 minute
d<-Extract_MSAP_epigenotypes ("MSAP_data.txt","Mix2", "MSAP_Mix2.txt",1,TRUE)

# Read transformed MSAP data and calculate population level descriptors
# Results are available as data.frame "p" and output to  "MSAP_Mix2_descr.txt"
p<-descriptive_parameters    ("MSAP_Mix2.txt","MSAP_Mix2_descr.txt")

#############################################################################
# Principle coordinates analysis
library(labdsv)                                  #load library for PCoA analysis
d.pco<-pco(dsvdis(d[,-c(1:2)],index="sorensen"),k=10) # perform PCoA

# simple PCoA plot
plot(d.pco)

# plot population IDs instead of symbols
plot(d.pco$points[,2]~d.pco$points[,1], type="n",
     xlab=paste("PCO1 (",format(d.pco$eig[1]/sum(d.pco$eig)*100,digits=3),"%)",
                sep=""),
     ylab=paste("PCO2 (",format(d.pco$eig[2]/sum(d.pco$eig)*100,digits=3),"%)",
                sep=""),
     main="Epigenetic variation of 3 pops")
text(d.pco$points[,1],d.pco$points[,2],d[,1],cex=0.6)

# simple barplot of overall Shannon diversity
barplot(p$Shannon_diversity,beside=T,names.arg=p$PopID,
          ylab="Mean Shannon diversity", xlab="Population",
          ylim=c(0,max(p$Shannon_diversity)*1.2), main="Mixed2 scoring")

# barplot of Shannon diversity of different types of epi-loci
barplot(t(rbind(p$u_Shannon_diversity,p$m_Shannon_diversity,
                p$h_Shannon_diversity)),
        beside=T,names.arg=p$PopID,
        ylab="Mean Shannon diversity", xlab="Population",
        ylim=c(0,max(cbind(p$u_Shannon_diversity,p$m_Shannon_diversity,
             p$h_Shannon_diversity))*1.4),
        legend.text= c("unmeth.","Me-CG","HMe-CCG"),args.legend=list(horiz=T))
#############################################################################
```

…. produces these figures









R code is available at: http://www.ufz.de/index.php?en=816

How to cite

If you use MSAP_calc_1_3.r, please cite Schulz, B., Eckstein R. L. and Durka W. (2013). Scoring and analysis of methylation sensitive amplification polymorphisms (MSAP) for epigenetic population studies. Molecular Ecology Resources **13**: 642-653.

References

Herrera CM, Bazaga P (2010) Epigenetic differentiation and relationship to adaptive genetic divergence in discrete populations of the violet *Viola cazorlensis*. *New Phytologist* **187**, 867-876.

Lira-Medeiros CF, Parisod C, Fernandes RA, *et al.* (2010) Epigenetic variation in mangrove plants occurring in contrasting natural environment. *Plos One* **5**, e10326.

Paun O, Bateman RM, Fay MF, *et al.* (2010) Stable epigenetic effects impact adaptation in allopolyploid orchids (*Dactylorhiza*: Orchidaceae). *Molecular Biology and Evolution* **27**, 2465-2473.

R Core Team (2012) *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Salmon A, Clotault J, Jenczewski E, Chable V, Manzanares-Dauleux MJ (2008) *Brassica oleracea* displays a high level of DNA methylation polymorphism. *Plant Science* **174**, 61-70.

Schulz B, Eckstein RL, Durka W (2013) Scoring and analysis of methylation-sensitive amplification polymorphisms for epigenetic population studies. *Molecular Ecology Resources* **13**, 642-653.

Schulz B, Eckstein RL, Durka W (2014) Epigenetic variation reflects dynamic habitat conditions in a rare floodplain herb. *Molecular Ecology* **23**, 3523–3537.

Vergeer P, Wagemaker N, Ouborg NJ (2012) Evidence for an epigenetic role in inbreeding depression. *Biology Letters* **8**, 798-801.

Papers citing MSAP_calc

Alonso C, Perez R, Bazaga P, Medrano M, Herrera CM (2016). MSAP markers and global cytosine methylation in plants: a literature survey and comparative analysis for a wild-growing species. Molecular Ecology Resources 16:80-90 doi: 10.1111/1755-0998.12426.

Alsdurf J, Anderson C, Siemens DH (2016). Epigenetics of drought-induced trans-generational plasticity: consequences for range limit development. Aob Plants 8 doi: 10.1093/aobpla/plv146.

Avramidou EV, Doulis AG, Aravanopoulos FA (2015a). Determination of epigenetic and genetic inheritance, and estimation of genome DNA methylation in a full-sib family of Cupressus sempervirens L. Gene doi: http://dx.doi.org/10.1016/j.gene.2015.02.068.

Avramidou EV, Ganopoulos IV, Doulis AG, Tsaftaris AS, Aravanopoulos FA (2015b). Beyond population genetics: natural epigenetic variation in wild cherry (Prunus avium). Tree Genetics & Genomes 11 doi: 10.1007/s11295-015-0921-7.

Baldanzi S, Watson R, McQuaid CD, Gouws G, Porri F (2017). Epigenetic variation among natural populations of the South African sandhopper Talorchestia capensis. Evolutionary Ecology 31:77-91 doi: 10.1007/s10682-016-9877-9.

Diao S, Wang Y, Ding C, Chang Y, Liang L, Gao Y, Zhang B, Su X (2016). No consistent daily variation in DNA methylation detected in Populus nigra leaves by methylation-sensitive amplification polymorphism analysis. Journal of Forestry Research:1-8 doi: 10.1007/s11676-016-0357-4.

Herrera CM, Bazaga P (2016a). Genetic and epigenetic divergence between disturbed and undisturbed subpopulations of a Mediterranean shrub: a 20-year field experiment. Ecology and Evolution 6:3832-3847 doi: 10.1002/ece3.2161.

Herrera CM, Medrano M, Bazaga P (2016b). Comparative spatial genetics and epigenetics of plant populations: heuristic value and a proof of concept. Molecular Ecology 25:1653-1664 doi: 10.1111/mec.13576.

Hofmann GE (2017). Ecological Epigenetics in Marine Metazoans. Frontiers in Marine Science 4 doi: 10.3389/fmars.2017.00004.

Inácio V, Barros PM, Costa A, Roussado C, Gonçalves E, Costa R, Graça J, Oliveira MM, Morais-Cecílio L (2017). Differential DNA Methylation Patterns Are Related to Phellogen Origin and Quality of Quercus suber Cork. Plos One 12:e0169018 doi: 10.1371/journal.pone.0169018.

Kellenberger RT, Schluter PM, Schiestl FP (2016). Herbivore-Induced DNA Demethylation Changes Floral Signalling and Attractiveness to Pollinators in Brassica rapa. Plos One 11 doi: 10.1371/journal.pone.0166646.

Kesäniemi JE, Heikkinen L, Knott KE (2016). DNA Methylation and Potential for Epigenetic Regulation in Pygospio elegans. Plos One 11 doi: 10.1371/journal.pone.0151863.

Kilaso M, Chapman NC, Remnant EJ, Oldroyd BP, Chanchao C (2016). No evidence that DNA methylation is associated with the regulation of fertility in the adult honey bee Apis mellifera (Hymenoptera: Apidae) worker ovary. Austral Entomology:n/a-n/a doi: 10.1111/aen.12218.

Machczynska J, Orlowska R, Zimny J, Bednarek PT (2014). Extended metAFLP approach in studies of tissue culture induced variation (TCIV) in triticale. Molecular Breeding 34:845-854 doi: 10.1007/s11032-014-0079-2.

Medrano M, Herrera CM, Bazaga P (2014). Epigenetic variation predicts regional and local intraspecific functional diversity in a perennial herb. Molecular Ecology 23:4926-4938 doi: 10.1111/mec.12911.

Mensaert K, Denil S, Trooskens G, Van Criekinge W, Thas O, De Meyer T (2014). Next-generation technologies and data analytical approaches for epigenomics. Environmental and Molecular Mutagenesis 55:155-170 doi: 10.1002/em.21841.

Nicotra AB, Segal DL, Hoyle GL, Schrey AW, Verhoeven KJF, Richards CL (2015). Adaptive plasticity and epigenetic variation in response to warming in an Alpine plant. Ecology and Evolution 5:634-647 doi: 10.1002/ece3.1329.

Piertney SB (2016). High-Throughput DNA Sequencing and the Next Generation of Molecular Markers in Wildlife Research. In Mateo R, Arroyo B, Garcia TJ (eds) Current Trends in Wildlife Research, Cham: Springer International Publishing, 201-223.

Preite V, Snoek LB, Oplaat C, Biere A, van der Putten WH, Verhoeven KJF (2015). The epigenetic footprint of poleward range-expanding plants in apomictic dandelions. Molecular Ecology 24:4406-4418 doi: 10.1111/mec.13329.

Rosellini D, Ferradini N, Allegrucci S, Capomaccio S, Zago ED, Leonetti P, Balech B, Aversano R, Carputo D, Reale L, Veronesi F (2016). Sexual Polyploidization in Medicago sativa L.: Impact on the Phenotype, Gene Transcription, and Genome Methylation. G3-Genes Genomes Genetics 6:925-938 doi: 10.1534/g3.115.026021.
Schulz B, Eckstein RL, Durka W (2014). Epigenetic variation reflects dynamic habitat conditions in a rare floodplain herb. Molecular Ecology 23:3523–3537 doi: DOI: 10.1111/mec.12835.
Smith TA, Martin MD, Nguyen M, Mendelson TC (2016). Epigenetic divergence as a potential first step in darter speciation. Molecular Ecology 25:1883-1894 doi: 10.1111/mec.13561.
Sun Y, Hou R, Fu X, Sun C, Wang S, Wang C, Li N, Zhang L, Bao Z (2014). Genome-Wide Analysis of DNA Methylation in Five Tissues of Zhikong Scallop, Chlamys farreri. Plos One 9 doi: 10.1371/journal.pone.0086232.
Wenzel MA, Piertney SB (2014). Fine-scale population epigenetic structure in relation to gastrointestinal parasite load in red grouse (Lagopus lagopus scotica). Molecular Ecology 23:4256-4273 doi: 10.1111/mec.12833.
Xia H, Huang WX, Xiong J, Tao T, Zheng XG, Wei HB, Yue YX, Chen L, Luo LJ (2016). Adaptive Epigenetic Differentiation between Upland and Lowland Rice Ecotypes Revealed by Methylation-Sensitive Amplified Polymorphism. Plos One 11 doi: 10.1371/journal.pone.0157810.
Xu JH, Tanino KK, Robinson SJ (2016). Stable Epigenetic Variants Selected from an Induced Hypomethylated Fragaria vesca Population. Frontiers in Plant Science 7 doi: 10.3389/fpls.2016.01768.
Zhu XL, Dong LM, Jiang LB, Li H, Sun LD, Zhang H, Yu WW, Liu HK, Dai WS, Zeng YR, Wu RL (2016). Constructing a linkage-linkage disequilibrium map using dominant-segregating markers. DNA Research 23:1-10 doi: 10.1093/dnares/dsv031

Upgrades

From MSAP_calc 1.0 to MSAP_calc 1.1:

- Identical sample-IDs are allowed in different populations. This lead to meaningless results in v. 1.0
- Bug corrected caused when only 1 subepilocus had to be extracted.

From MSAP_calc 1.1 to MSAP_calc 1.2:

- Another bug for identical sample-IDs in different populations. Bugs in MSAP_calc_session.R

From MSAP_calc 1.2 to MSAP_calc 1.3:

- Correcting bugs in previous versions for identical sample-IDs in different populations.