Appendix S2: Using AveDissR.pdf as part of the AveDissR.rar package

Instructions for using AveDissR for assessing genetic distinctness and genetic redundancy

There are six steps to follow for using AveDissR. Please read these steps carefully before running AveDissR.r

1. Install R, if not available in the computer or Unix-like server

Follow the installation instructions to install R (https://www.r-project.org/) in Windows or Linux server; or R Studio (https://www.rstudio.com/) in Windows. In a Linux server, you may need to set the path to R for use.

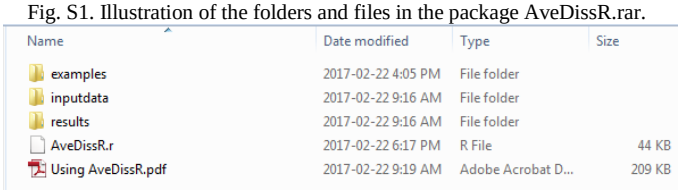
2. Install five R packages required for using AveDissR

>install.packages("reshape2", dependencies = TRUE)
>install.packages("data.table", dependencies = TRUE)
>install.packages("vegan", dependencies = TRUE)
>install.packages("foreach",dependencies = TRUE)
>install.packages("doParallel",dependencies = TRUE)

3. Download AveDissR.rar from the journal website

The package AveDissR.rar has one R file AveDissR.r, the instruction file *Using AveDissR.pdf*, and three subfolders named *examples*, *inputdata*, and *result*s. Unzip AveDissR.rar into a directory called *AveDissR* in either your personal computer or a Linux server, and it will display as shown in Fig. S1.

Fig. S1. Illustration of the folders and files in the package AveDissR.rar.

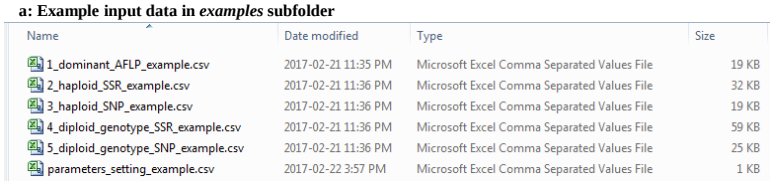


Name	Date modified	Type	Size
examples	2017-02-22 4:05 PM	File folder	
inputdata	2017-02-22 9:16 AM	File folder	
results	2017-02-22 9:16 AM	File folder	
AveDissR.r	2017-02-22 6:17 PM	R File	44 KB
Using AveDissR.pdf	2017-02-22 9:19 AM	Adobe Acrobat D...	209 KB

The subfolder *inputdata* is used to place the assessed marker data and its parameters setting, respectively. The subfolder *results* where the output files are generated from the analysis. The subfolder *examples* has one example file for parameter setting and five example marker data files, as shown in Fig. S2.

Fig. S2. Illustration of the example files in the subfolder *examples.*

a: Example input data in *examples* subfolder



Name	Date modified	Type	Size
1_dominant_AFLP_example.csv	2017-02-21 11:35 PM	Microsoft Excel Comma Separated Values File	19 KB
2_haploid_SSR_example.csv	2017-02-21 11:36 PM	Microsoft Excel Comma Separated Values File	32 KB
3_haploid_SNP_example.csv	2017-02-21 11:36 PM	Microsoft Excel Comma Separated Values File	19 KB
4_diploid_genotype_SSR_example.csv	2017-02-21 11:36 PM	Microsoft Excel Comma Separated Values File	59 KB
5_diploid_genotype_SNP_example.csv	2017-02-21 11:36 PM	Microsoft Excel Comma Separated Values File	25 KB
parameters_setting_example.csv	2017-02-22 3:57 PM	Microsoft Excel Comma Separated Values File	1 KB

b: parameters_setting_example.csv

	A	B	
1	Parameters	Value	Value setting description
2	Analysis_Type	1	1 for genetic distinctness analysis, 2 for gene
3	Marker_Type	1	1 for dominant marker; 2 or 3 for haploid SSR
4	SampleN	670	Total sample size
5	PopulationN	80	The number of populations
6	LociN	170	The number of loci
7	Missing_Label	NA	Missing value is defined as NA
8	stepwise selection	0.1	Outputs for stepwise selections can be adjus
9	specific selection	NA	Output for a specified proportion (0 to 1) of s

c: : 1_dominant_AFLP_example.csv

	A	B	C	D	E	F	G	H	I
1	Ind	Ind_001	Ind_002	Ind_003	Ind_004	Ind_005	Ind_006	Ind_007	Ind_008
2	Pop	pop01	pop01	pop01	pop01	pop01	pop01	pop02	pop02
3	loci_001	1	1	1	0	0	1	1	1
4	loci_002	0	1	1	0	0	0	1	0
5	loci_003	1	0	1	1	0	0	1	1
6	loci_004	0	0	NA	0	0	NA	1	1
7	loci_005	1	0	0	1	1	1	1	0
8	loci_006	1	1	1	0	1	NA	1	0
9	loci_007	1	NA	NA	1	1	0	1	0

d: 2_haploid_SSR_example.csv

	A	B	C	D	E	F	G	H	I
1	Ind	Ind_001	Ind_002	Ind_003	Ind_004	Ind_005	Ind_006	Ind_007	Ind_008
2	Pop	pop01	pop01	pop01	pop01	pop01	pop01	pop02	pop02
3	loc_001	223	NA	356	223	259	356	283	283
4	loc_002	223	178	223	356	239	178	259	259
5	loc_003	223	356	178	NA	239	178	NA	283
6	loc_004	223	283	NA	NA	259	239	289	178
7	loc_005	NA	NA	239	223	239	239	289	289
8	loc_006	NA	223	259	356	223	259	259	259
9	loc_007	178	259	223	259	223	356	259	NA

e: 3_haploid_SNP_example.csv

	A	B	C	D	E	F	G	H	I
1	Ind	Ind_001	Ind_002	Ind_003	Ind_004	Ind_005	Ind_006	Ind_007	Ind_008
2	Pop	pop01	pop01	pop01	pop01	pop01	pop01	pop02	pop02
3	loci_001	G	G	T	T	T	T	G	T
4	loci_002	G	C	C	G	C	C	NA	G
5	loci_003	C	A	NA	A	C	C	A	A
6	loci_004	C	A	C	C	C	NA	C	C
7	loci_005	G	T	G	G	T	G	T	T
8	loci_006	G	G	G	G	G	G	A	G
9	loci_007	T	T	A	T	A	T	A	A

f: 4_diploid_genotype_SSR_example.csv

	A	B	C	D	E	F	G	H	I
1	Ind	Ind_001	Ind_002	Ind_003	Ind_004	Ind_005	Ind_006	Ind_007	Ind_008
2	Pop	pop01	pop01	pop01	pop01	pop01	pop01	pop02	pop02
3	loci_001	121:121	121:385	121:202	385:385	121:236	202:236	233:385	121:233
4	loci_002	121:385	121:385	202:202	NA	236:236	202:385	121:256	121:236
5	loci_003	236:278	121:236	NA	278:278	236:278	236:278	233:233	202:236
6	loci_004	202:278	202:278	202:328	NA	236:328	202:202	121:256	NA
7	loci_005	121:385	121:236	328:385	236:385	328:385	121:236	202:236	202:233
8	loci_006	121:385	236:385	121:202	202:385	202:385	202:202	NA	121:202
9	loci_007	121:236	121:236	121:236	121:202	202:236	121:202	236:256	256:256

g: 5_diploid_genotype_SNP_example.csv

	A	B	C	D	E	F	G	H	I
1	Ind	Ind_001	Ind_002	Ind_003	Ind_004	Ind_005	Ind_006	Ind_007	Ind_008
2	Pop	pop01	pop01	pop01	pop01	pop01	pop01	pop01	pop01
3	loci_001	AC	AC	AC	AC	AC	CC	CC	CC
4	loci_002	NA	AA	AC	AA	AG	AG	AA	AC
5	loci_003	AA	AC	CC	AA	CC	NA	CC	AC
6	loci_004	CT	CC	CT	CT	TT	TT	CT	TT
7	loci_005	AC	AC	AC	NA	CC	CC	AC	CC
8	loci_006	NA	CG	CG	CC	CG	CG	CG	CG
9	loci_007	CC	CT	CT	CC	CC	CC	CT	TT

4. Prepare *marker* data and *parameters setting* data

Each analysis requires two input data sets as "parameters_setting.csv" and 1_myInputMarkerFile.csv (or .txt). Thus, one has to follow the formats shown in the subfolder *examples* to prepare these two files for each analysis. To assist the formatting, the following instructions are provided as below.

4.1 *marker* data format

The first row in the myInputMarkerFile.csv file is used for sample label, starting from the 2nd column. Each sample has a unique numerical identifier. The second row is used for population label, starting from the 2nd column with same prefix character and unique numerical label according to the total population size. In each population, all samples are grouped together. The third row onward is used for marker data. For each row of marker data, the first column is used to define the name of a locus, while the other columns are used for the marker data for specific samples. Pay specific attention to the marker data format and select one data format that matches with your marker data to prepare the input file. Note that all the missing values are defined as NA in each marker data file. Specific to haploid_SNP marker data, all the ambiguous nucleotides of K, V, H, D, B, N, M, R, W, S and Y should be treated as NA. The completed marker data needs to be saved as either as .csv (Comma delimited) or .txt (Tab delimted) file and can be named starting with the number prefixed with marker type such as 5_myInputMarkerFile.csv for diploid_genotype_SNP.csv. Also note the difference between .csv and .xlsx file extension.

4.2 *parameters_setting* format

Parameters_setting.csv has eight parameters to be defined in three columns. The first column is parameter name, the second column is used to set parameter value, and the third column is for parameter value description. The *Analysis_Type* in B2 has only two options (1 for genetic distinctness analysis and 2 for genetic redundancy). The *Marker_Type* in B3 has five options: 1 for AFLP or dominant marker; 2 and 3 for haploid SSR and SNP data, respectively; 4 for diploid genotype data of SSR with a separator of ":"between two alleles; and 5 for diploid genotype data of SNP without a separator format for two alleles. Following B4, B5 and B6 are for sample size (SampleN), population size (PopulationN) and the number of loci (LociN), respectively, according to the assessed marker data. Missing data is labelled as NA in B7. The last two parameters (stepwise selection and specific selection) are used to define the outputs of AMOVA and PCoA analyses. The outputs for stepwise selections can be adjusted by specifying a starting proportional increment (or size of each iterative step) from 0 to 1 in B8 (default=0.1 with 9 steps; changing to 0.13 would have 7 steps from 0.13 to 0.91). NA=no stepwise selection. Note PCoA outputs are provided only for the first three proportional selections. For specific proportional selection, the output for a specified proportion (0 to 1) of selection can be made (e.g, 0.23) in B9. NA=no specific selection. Note that the AveDissR function only recognizes the parameter file name as "parameters_setting.csv" in subfolder *inputdata*.

5. Place two input data files into *inputdata* subfolder

Place *parameters_setting.csv* and myInputMarkerFile.csv to the *inputdata* subfolder. Only two input data files are allowed in the *inputdata* subfolder. Also, the names of the original subfolders unzipped from AveDissR.rar are not allowed to change.

6. Run AveDissR.r

Before running AveDissR.r, you need to make sure that five R packages of "reshape2", "data.table", "vegan", "foreach", and "doParallel" are successfully installed.

6.1 Run AveDissR.r without opening of R console

In Linux, you need to go to the directory of AveDissR, and type: R CMD BATCH AveDissR.r &, assuming R is accessible in the AveDissR directory. If R path is not set up, type: yourPathToR/R CMD BATCH AveDissR.r &.

In Windows CMD, you need to go to the directory of AveDissR in your computer, and type: "yourPathToR\bin\x64\Rscripts.exe" AveDissR.r. For example, type: "C:\Program Files\R-3.2.3\bin\x64\Rscripts.exe" AveDissR.r.

6.2 Run AveDissR.r within R console

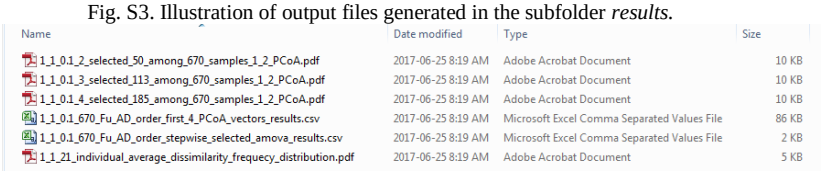
In Linux with the opening R console (or by typing R in any Linux directory), you need to set the working directory to where AveDissR is by typing: setwd ("yourPath/AveDissR"). For example, setwd ("/home/fuy/allelicCount/AveDissR"). If it works, then type: source ("AveDissR.r").

In R console in Windows or R console under Rstudio, you need to set the working directory to where AveDissR is by typing: setwd ("yourPath/AveDissR"). For example, setwd ("d:/Rdistinct/AveDissR"). If it works, then type: source ("AveDissR.r").

6.3 Get AveDissR results in *results* subfolder.

If AveDissR runs successfully, a set of output files for each analysis will be generated in the *results* subfolder. With default stepwise selection, six files are generated to display the analysis results. For specific selection, four files are produced to display the results. Following in Fig. S3 are the example output files generated in the *results* subfolder from the stepwise selection for genetic distinctness analysis.

Fig. S3. Illustration of output files generated in the subfolder *results*.



7 Note AveDissR.r was successfully tested in R version 3.2.3 Platform: x86_64-w64-mingw32/x64 (64-bit) (Windows) and Platform: x86_64-redhat-linux-gnu (64-bit) (Linux). For example, running the soybean data set of 4,500 accessions x 6,000 SNPs (including missing data) with 40-cores parallel computing in the Linux server lasted two hours and 48 minutes. AveDissR.r has not yet been tested in an R environment under the computer operating system Mac OS X, but theoretically can be similarly applied.