AveDissR: An R Function for Assessing Genetic Distinctness and Genetic Redundancy

Authors: Yang, Mo-Hua, and Fu, Yong-Bi

Source: Applications in Plant Sciences, 5(7)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1700018

Software Note

AveDissR: An R function for assessing genetic distinctness and genetic redundancy[1]

Mo-Hua Yang[2,3] and Yong-Bi Fu[2,4]

[2]Plant Gene Resources of Canada, Saskatoon Research and Development Centre, Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon S7N 0X2, Canada; and [3]College of Forestry, Central South University of Forestry and Technology, 498 South Shaoshan Road, Changsha, Hunan 410004, People's Republic of China

- *Premise of the study:* Assessing genetic distinctness or redundancy is an important part of plant germplasm characterization. We previously introduced a new marker-based approach using the average dissimilarity of an accession to assess genetic distinctness or redundancy. However, this approach has not been widely applied, largely due to the lack of software to integrate separate analyses involving dissimilarity calculation, analysis of molecular variance, and principal coordinates analysis.
- *Methods and Results:* An R function, AveDissR, was developed to integrate three separate analyses into one package for assessing genetic distinctness or redundancy. It can analyze large data sets of dominant or codominant markers such as amplified fragment length polymorphisms (AFLPs), simple sequence repeats (SSRs), or single-nucleotide polymorphisms (SNPs), generate a useful set of output files for germplasm assessment, and run in an R environment on any computer platform.
- *Conclusions:* AveDissR can make the assessment of genetic distinctness or redundancy in plant germplasm more feasible and useful.

Key words: AMOVA; average dissimilarity; genetic distinctness; genetic redundancy; germplasm characterization; PCoA.

Concerted conservation efforts over the past 50 yr have collected more than seven million plant germplasm accessions of more than 16,500 plant species that are currently conserved in 1750 genebanks worldwide (Engels, 2004; FAO, 2010). Management and utilization of these germplasm collections is a challenging mission (Engels and Visser, 2003), as large efforts are required to characterize these germplasm collections and only two million accessions are estimated to be unique (FAO, 2010). More core subsets need to be developed from large collections for germplasm screening of specific traits of interest (Frankel, 1984; Brown, 1989; van Hintum et al., 2000). Thus, assessing genetic distinctness or redundancy has become an important part of germplasm characterization (Waycott and Fort, 1994; Virk et al., 1995; Phippen et al., 1997; Chavarriaga-Aguirre et al., 1999; Dean et al., 1999; Karp, 2002; Fu, 2006; Kisha and Cramer, 2011; Motilal et al., 2013). Identification of genetically distinct germplasm is useful for the establishment of core subsets in a germplasm collection, and assessment of genetically redundant germplasm can help to validate accession duplication. Recent technical advances in the development of molecular markers will make molecular characterizations of plant germplasm more feasible than before (Fu and Peterson, 2012; Peterson et al., 2014; Song et al., 2015).

To facilitate germplasm characterization, we previously introduced a new marker-based approach using the average dissimilarity (AD) of an accession to assess genetic distinctness or genetic redundancy in a plant germplasm collection (Fu, 2006). The AD approach is based on the acquired molecular characterization data, generates the average dissimilarity of an accession against the remaining assayed accessions, and provides a means to identify genetically distinct or redundant germplasm. The approach has been well cited in the scientific literature, but it has not been applied as widely as hoped for. We reasoned that the lack of integrated computer software to streamline separate analyses involving dissimilarity calculation, analysis of molecular variance (AMOVA; Excoffier et al., 1992), and principal coordinates analysis (PCoA) is the major factor for its limited application.

Here we present an R function (R Core Team, 2016; https://www.r-project.org), called AveDissR, which integrates three separate analyses of the AD approach into one package for identifying genetically distinct or redundant germplasm. This R function considers five different data types of dominant or codominant molecular markers such as amplified fragment length polymorphisms (AFLPs), simple sequence repeats (SSRs), and single-nucleotide polymorphisms (SNPs); generates a useful set of output files for germplasm assessment; and can be run in R environments under different computer platforms. This newly developed R function will make the assessment of genetic

[1] Manuscript received 1 March 2017; revision accepted 29 May 2017.

The authors thank Dr. Qijin Song (U.S. Department of Agriculture–Agricultural Research Service) for his assistance in acquisition of soybean single-nucleotide polymorphism data for the analysis and two anonymous reviewers for their helpful comments on the manuscript. This research was financially supported by an Agriculture and Agri-Food Canada A-Base research project to Y.B.F., the National Natural Science Foundation of China (no. 31670678), and a China Scholarship Council Postdoctoral Abroad Grant to M-H.Y.

[4]Author for correspondence: yong-bi.fu@agr.gc.ca

doi:10.3732/apps.1700018

distinctness and/or genetic redundancy in plant germplasm characterizations more feasible and useful.

METHODS AND RESULTS

The average dissimilarity approach—The original approach introduced by Fu (2006) is conceptually simple and has three major components. The first component is to calculate AD for each accession. In a typical marker-based characterization of self-fertile plant germplasm, an accession is usually represented by a single plant; *n* accessions are selected from a collection to represent *p* countries (or regions) of origin; and these accessions are assayed with molecular markers of *m* loci. Thus, a given accession can form *n* − 1 pairwise pairs with the remaining assayed accessions. For dominant markers (e.g., AFLP), the pairwise similarity S_{ij} between accessions *i* and *j* can be calculated following the simple matching coefficient of Sokal and Michener (1958) as:

$$S_{ij}=\frac{a}{a+b}, \qquad [1]$$

where *a* is the number of matched AFLP bands (or genotypes) between accessions *i* and *j* across nonmissing loci m_{nm} ($m_{nm} \le m$), and *b* is the number of mismatched bands (or genotypes) between *i* and *j* across m_{nm} loci. Note that any loci with one or two accessions having no genotype values are not counted, and consequently $m_{nm} \le m$ in the pair. The dissimilarity for such pair D_{ij} is defined as $1 - S_{ij}$ [$= b/(a + b)$]. The AD value for accession *i* is obtained by averaging all *n* − 1 pairwise accession dissimilarities D_{ij} as:

$$AD_i=\frac{1}{n-1}\sum_{j=1}^{n-1}(1-S_{ij}). \qquad [2]$$

The higher the AD value, the more genetically distinct the accession is in the collection. The lower the AD value, the more genetically redundant the accession is in the collection. Thus, ranking the AD values of all *n* accessions provides a means of identifying the genetically most distinct and the genetically most redundant accessions. However, such ranking does not provide a threshold that can be used to define a group of genetically distinct or redundant accessions. Thus, the second component of the original approach was configured with the hope to identify a group of genetically redundant or distinct accessions. Fu (2006) explored a means of iteration through the selection of accessions with the lowest AD values and AMOVA to generate phi-statistics for the assessment of the impact of removing tentatively redundant accessions on the change of genetic differentiation present in the remaining assayed accessions. The redundant group of accessions selected with the lowest AD values will be defined when the genetic differentiation measured by phi-statistics in the remaining set of germplasm is largely maintained (or is within a nominal tolerance of 5% change in genetic differentiation from the originally assayed accessions). Similarly, to identify a group of the genetically most distinct accessions, the same means of iteration is applied through selection of accessions with the highest AD values and AMOVA to assess the change of genetic differentiation in the selected accessions. The distinct group of accessions selected with the highest AD values will be defined when the genetic differentiation inferred in the selected accessions is compatible with that present in the original assayed accessions (that is, within a nominal tolerance of 5% change in genetic differentiation). However, the iterative AMOVA may not always be informative when the representative groups are small and the inferred genetic differentiation may greatly deviate or fluctuate from the original one. Consequently, we introduced the third component to assess the representativeness of the accessions from AD-based identification relative to the whole set of samples through a PCoA plot of the genetic associations between the identified group and the whole set. This is done based on the dissimilarity matrix of all pairwise accessions, and the PCoA plot is generated with separate labels for the accessions of the identified group and the whole set.

To assist the application of the AD approach to analyze the flax collection of 2727 accessions genotyped with 149 polymorphic RAPD markers, we developed a SAS routine in 2006 (SAS Institute, 2004) to integrate the first two components of the approach, but not the third component of visual assessment over a PCoA plot. The latter was performed separately using NTSYS-pc software (Rohlf, 1997; Fu, 2006). Clearly, a set of computer software to streamline separate analyses is desirable for the wider application of the AD approach.

The R function: AveDissR—To facilitate the identification of genetically distinct or redundant germplasm, we developed an R function, called AveDissR, to integrate dissimilarity calculation, AMOVA, and PCoA in the original AD approach into one package. To widen its application with different markers, we also considered not only dominant or haploid marker data in the original AD approach, but also expanded it to deal with different codominant marker data. For a diploid codominant genotype marker data with *m* loci assayed for *n* accessions, the pairwise accession dissimilarity D_{ij} is calculated as the proportion of marker loci having different marker genotypes, which is analogous to the Hamming distance (Hamming, 1950; Wang et al., 2015) as below:

$$D_{ij}=\frac{b_{ij}}{m_{nm}}, \qquad [3]$$

where b_{ij} is the number of dismatched genotypes over nonmissing loci m_{nm} ($m_{nm} \le m$) between the paired accessions *i* and *j* (*j* = 1.. *n* − 1). For example, diploid SSR genotypes 242 : 266 and 242 : 254 (or 242 : 242) for the paired accessions are counted as dismatched at the SSR locus with three alleles of size: 242, 254, and 266, while the two genotypes 254 : 266 and 254 : 266 are matched. The AD value for accession *i* is calculated as the same as for dominant marker data by averaging all *n* − 1 pairwise accession dissimilarities D_{ij}:

$$AD_i=\frac{1}{n-1}\sum_{j=1}^{n-1}D_{ij}. \qquad [4]$$

Accordingly, its AMOVA component also accommodates both dominant and codominant marker data for analysis of molecular variance following the established methods (Excoffier et al., 1992; Peakall and Smouse, 2006, 2012). The PCoA component is dependent only on a pairwise accession dissimilarity matrix, and the matrix is calculated as mentioned above for both dominant and codominant marker data.

Specifically, AveDissR was configured and developed to analyze five different types of marker data: (1) AFLP or dominant marker; (2 and 3) haploid SSR and SNP data, respectively; (4) diploid genotype data of SSR or SNP with a separator of ":" between two alleles; and (5) diploid genotype data of SNP without a separator format for two alleles. The R function was written with five major steps: (1) data input and dissimilarity matrix calculation, (2) generating AD values for all accessions, (3) performing PCoA, (4) conducting iterative AMOVA, and (5) PCoA display and result outputs. To achieve these tasks, AveDissR utiltizes five R packages ("reshape2", "data.table", "vegan", "foreach", and "doParallel") and is equipped with several custom R functions to transform input data using "data.table" and "reshape2" utilities, calculate pairwise dissimilarity matrix and AD values, conduct two separate AMOVAs for dominant or codominant markers, perform PCoA vectors using "vegan" cmdscale, and generate PCoA plots and outputs. The packages "foreach" and "doParallel" were applied to use about half of existing processors or cores in a computing platform to speed up pairwise calculations. Depending on the analysis objective (i.e., to identify genetically distinct or genetically redundant germplasm), several output files are generated to obtain AD values and PCoA vectors, iterative AMOVA results, and PCoA plots.

This R function has several advantages over the original AD analysis with the developed SAS routine (Fu, 2006). First, it is a streamlined package that allows for one integrated analysis to generate AD value for each accession, as well as to have the iterative AMOVA and PCoA outputs to facilitate germplasm grouping. Second, it can analyze germplasm assayed with different marker types, making it easier to apply to a wide variety of problems in germplasm characterization. Third, the R function can take advantage of parallel computing to handle large data sets, particularly in operating systems Mac OS X or Linux. For example, running 4500 soybean (*Glycine* spp.) accessions genotyped at 6000 SNPs (including missing values) with 40-core parallel computing in a Linux server lasted 2 h 48 min, while the serial execution with the same data set consumed 43.92 h. Fourth, it is more accessible than the SAS routine (Fu, 2006), as R is not only a widely used programming language and software environment for statistical computing and graphics, but also freely available under the GNU General Public License. AveDissR is packaged as *AveDissR.rar* and is freely available for use with different computing systems under Windows, Mac OS, or Linux. The R script, user instruction file, and example data and output files are available on Figshare (http://dx.doi.org/10.6084/m9.figshare.5082451; Yang and Fu, 2017). Its use 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).

***Using AveDissR*—**The package *AveDissR.rar* (Yang and Fu, 2017: Appendix S1) has one R script file *AveDissR.r*, the user instruction file *Using AveDissR.pdf* (Yang and Fu, 2017: Appendix S2), and three subfolders named *examples*, *inputdata*, and *results* (Appendix S2: Fig. S1). The subfolder *examples* has five example data files and one *parameters_setting.example.csv* file (Appendix S2: Fig. S2a). The subfolder *inputdata* is set up to place marker data and *parameters_setting.csv* files for an analysis. The subfolder *results* is where the output files are generated from the analysis (Appendix S2: Fig. S3). *Using AveDissR.pdf* provides detailed instructions for using *AveDissR.r* in R environments under different computer settings (Appendix S2). There are several major steps to follow. They include (1) install R, if not available in computer or Unix-like server; (2) install five R packages required for using AveDissR; (3) download *AveDissR.rar* from the Figshare website (Yang and Fu, 2017); (4) prepare marker data and parameters_setting files for analysis; (5) place the marker data and parameters_setting files into the *inputdata* subfolder; and (6) run *AveDissR.r* either without opening R console or within R console, as shown in Table 1. Attention should be paid to prepare marker data and parameters_setting. The R function is capable of analyzing any of the five marker data types, but each marker data type has its unique format (Appendix S2: Fig. S2c–g). The marker data file should be prepared following one of the five example data file formats. The R function uses *fread* function to input data file with a *.csv* (comma delimited) or *.txt* (tab delimited) format. The parameters_setting file is used to instruct the AveDissR analysis (Appendix S2: Fig. S2b). Seven sets of parameters need to be defined before an analysis can start. The *parameters_setting.example.csv* file in the *examples* subfolder (Appendix S2: Fig. S2b) shows the description of, and the usage for, each parameter. More attention should be paid to the parameters *stepwise selection* and *specific selection*, as both settings determine the iterative AMOVA and PCoA outputs to assess the changes in genetic differentiation of selected or removed germplasm to facilitate the grouping of germplasm with higher or lower AD values. *Stepwise selection* allows users to select the starting proportion (or the size of the step) for iteration up to 1 (100% germplasm). If one decides to select only one specific proportion of germplasm (e.g., 26%), *stepwise selection* can be silenced with NA and *specific selection* can be defined as 0.26.

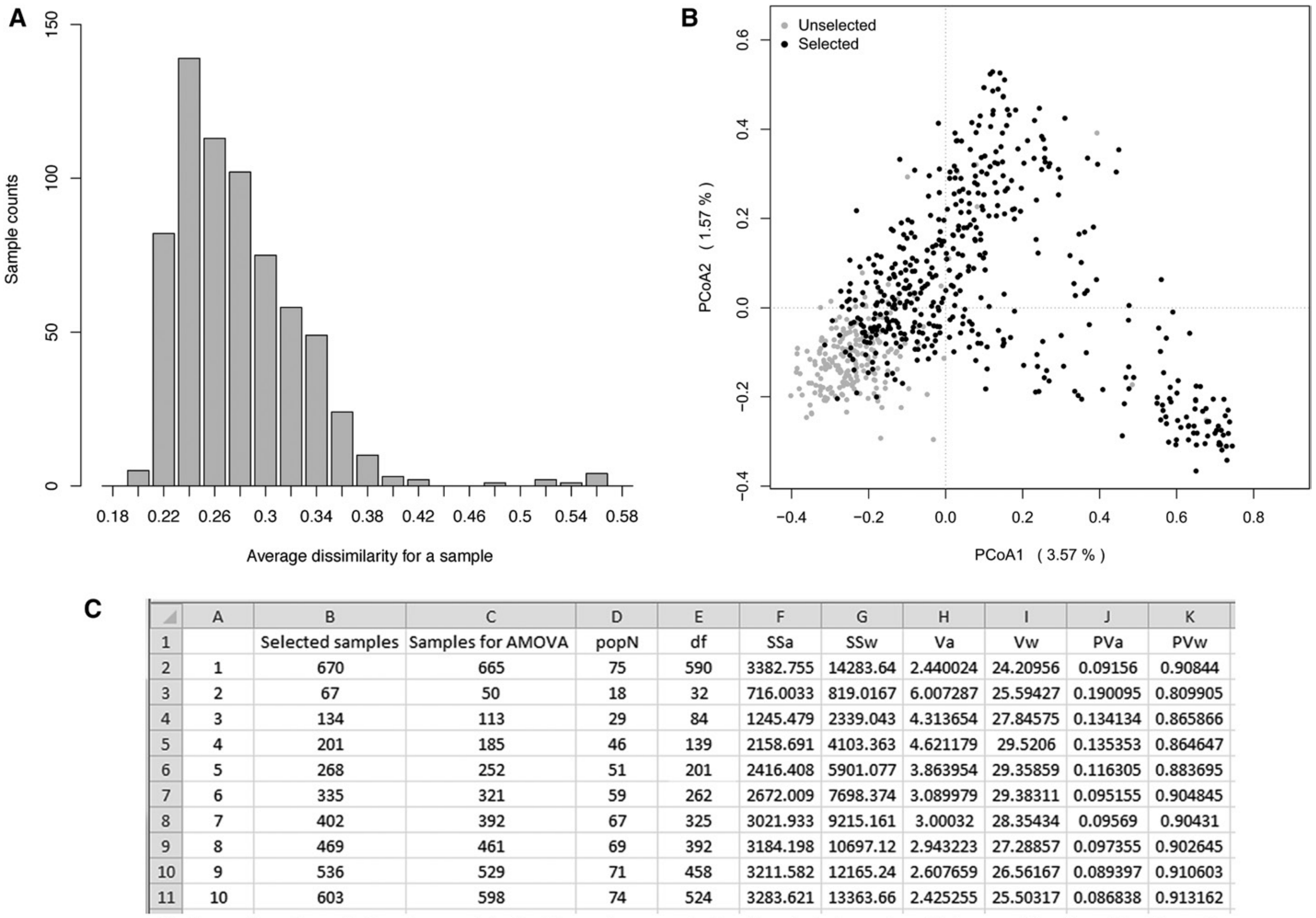
***Illustration of its usage*—**AveDissR can be applied to assess genetic distinctness or genetic redundancy separately or jointly, depending on the objectives of a plant germplasm characterization. Here we illustrate its usage to assess genetic distinctness and genetic redundancy separately, each with a published data set. Two illustrative data sets are available upon request from the corresponding author. The first data set selected is the AFLP data of 670 cultivated hexaploid oat (*Avena sativa*) accessions (Fu et al., 2005). The assayed accessions were selected from a world collection of 11,622 cultivated hexaploid oat accessions to represent 79 countries and one group of uncertain origin, and they were genotyped with 170 AFLP markers by five AFLP primer pairs. Previous AFLP analysis showed this set of oat germplasm was genetically diverse and distinct (Fu et al., 2005). Here we reanalyzed the data set using AveDissR to assess their genetic distinctness. We reformatted the data following *1_dominant_AFLP_example.csv* (Appendix S2: Fig. S2c) and defined the parameters_setting file as: *Analysis_Type* = 1, *Marker_Type* = 1, *SampleN* = 670, *PopulationN* = 80, *LociN* = 170, *Missing_Label* = NA, *stepwise selection* = 0.1, and *specific selection* = NA. Running the data in the mentioned Linux server lasted 3 min. Six original output files were generated: one bar plot for AD distribution, three PCoA plots for selected 50, 113, or 185 accessions, one .csv file for AD output and PCoA vectors, and one .csv file for iterative AMOVA output. To illustrate, three sets of the outputs are shown in Fig. 1: the distribution of average dissimilarity values in 670 accessions, the PCoA plot for the associations of 461 selected and 670 original accessions, and the AMOVA outputs showing the changes in genetic differentiation with the increased selection of accessions with the highest AD values. The AD values for these 670 oat accessions ranged widely from 0.214 to 0.575 with a mean of 0.289 (Fig. 1A). Selecting 469 out of 670 oat accessions with the highest AD values as the genetically distinct group could largely maintain the extent of genetic differentiation in the assayed 670 oat accessions, as shown with percentage of variation among populations (PVa) values of 0.09156 in the first step and 0.097355 in the eighth step with the selection of 469 samples, respectively (Fig. 1C). The PCoA plot (Fig. 1B) showed the representativeness of the 461 (out of 469 selected) samples (in the eighth step) used for AMOVA relative to the original 670 accessions, based on the revealed genetic associations. One could rerun *AveDissR.r* with different parameters_setting files by adjusting *stepwise selection* values to select a tentative group of genetically distinct accessions, as suggested by Fu (2006) with a nominal tolerance of 5% change in genetic differentiation from the originally assayed accessions. When the tentative distinct group is selected, *AveDissR.r* can be rerun with *specific selection* value (equal to the size of the selected group over the assayed accessions) to generate the final PCoA plot for assessment. Note that Fig. 1B was obtained after rerunning *AveDissR.r* with *specific selection* = 0.7 (= 469/670) and that the number of samples used for AMOVA may be smaller than the selected samples, as AMOVA removed those populations (or countries in this case) that had only a single sample (Fig. 1C).

The second data set was selected from the soybean data generated by Song et al. (2015) to illustrate its use in assessment of genetic redundancy. The original soybean data consists of 18,480 domesticated soybean and 1168 wild soybean accessions that were genotyped with 42,509 polymorphic SNP markers through the SoySNP50K Illumina Infinium II BeadChip (Illumina, San Diego, California, USA). The assayed soybean accessions represented germplasm introduced from 84 countries or developed in the United States. Here we reanalyzed part of the original soybean data set using AveDissR to illustrate the outputs for germplasm redundancy assessment. We extracted 1828 accessions from 18,480 *Glycine max* accessions with respect to country of origin and randomly selected 5000 (from 42,509) SNP markers to form a new soybean data set representing 21 countries. We reformatted the data following *3_haploid_SNP_example.csv* (Appendix S2: Fig. S2e) and defined the parameters_setting file as: *Analysis_Type* = 2, *Marker_Type* = 3, *SampleN* = 1828, *PopulationN* = 21, *LociN* = 5000, *Missing_Label* = NA, *stepwise selection* = 0.1, and *specific selection* = NA. Running the data in the Linux server lasted roughly 22 min. Six original output files were generated: one bar plot for AD distribution; three PCoA plots for selected 183, 366, or 549 accessions; one .csv file for AD output and PCoA vectors; and one .csv file for iterative AMOVA output. To illustrate, three sets of the outputs are shown in Fig. 2: the distribution of average dissimilarity values in 1828 accessions, the PCoA plot for the associations of 549 selected and 1828 original accessions, and the AMOVA outputs showing the changes in genetic differentiation with the increased removal of accessions with the lowest AD values. The AD values for these soybean accessions ranged widely from 0.28 to 0.487 with a

Table 1. Commands used to run *AveDissR.r* in the computer operating systems Windows or Linux and its working environment.

System[a]	R console	In the AveDissR folder	Command
Windows	Yes	Yes	> source (“AveDissR.r”)
Windows	Yes	No	> setwd (“yourPath\AveDissR”) > source (“AveDissR.r”)
Windows CMD	No	Yes	> “YourInstalledRpath\bin\x64\Rscript.exe” AveDissR.r
Windows CMD	No	No	> cd yourPathTo the AveDissR directory > “YourInstalledRpath\bin\x64\Rscript.exe” AveDissR.r
Linux	Yes	Yes	> source (“AveDissR.r”)
Linux	Yes	No	> setwd (“yourPath/AveDissR”) > source (“AveDissR.r”)
Linux	No	Yes	> R CMD BATCH AveDissR.r &
Linux	No	No	> cd yourPathTo the AveDissR directory > R CMD BATCH AveDissR.r &

[a] *AveDissR.r* has not yet been tested in an R environment under the operating system Mac OS X, but theoretically can be similarly applied.



	A	B	C	D	E	F	G	H	I	J	K
1		Selected samples	Samples for AMOVA	popN	df	SSa	SSw	Va	Vw	PVa	PVw
2	1	670	665	75	590	3382.755	14283.64	2.440024	24.20956	0.09156	0.90844
3	2	67	50	18	32	716.0033	819.0167	6.007287	25.59427	0.190095	0.809905
4	3	134	113	29	84	1245.479	2339.043	4.313654	27.84575	0.134134	0.865866
5	4	201	185	46	139	2158.691	4103.363	4.621179	29.5206	0.135353	0.864647
6	5	268	252	51	201	2416.408	5901.077	3.863954	29.35859	0.116305	0.883695
7	6	335	321	59	262	2672.009	7698.374	3.089979	29.38311	0.095155	0.904845
8	7	402	392	67	325	3021.933	9215.161	3.00032	28.35434	0.09569	0.90431
9	8	469	461	69	392	3184.198	10697.12	2.943223	27.28857	0.097355	0.902645
10	9	536	529	71	458	3211.582	12165.24	2.607659	26.56167	0.089397	0.910603
11	10	603	598	74	524	3283.621	13363.66	2.425255	25.50317	0.086838	0.913162

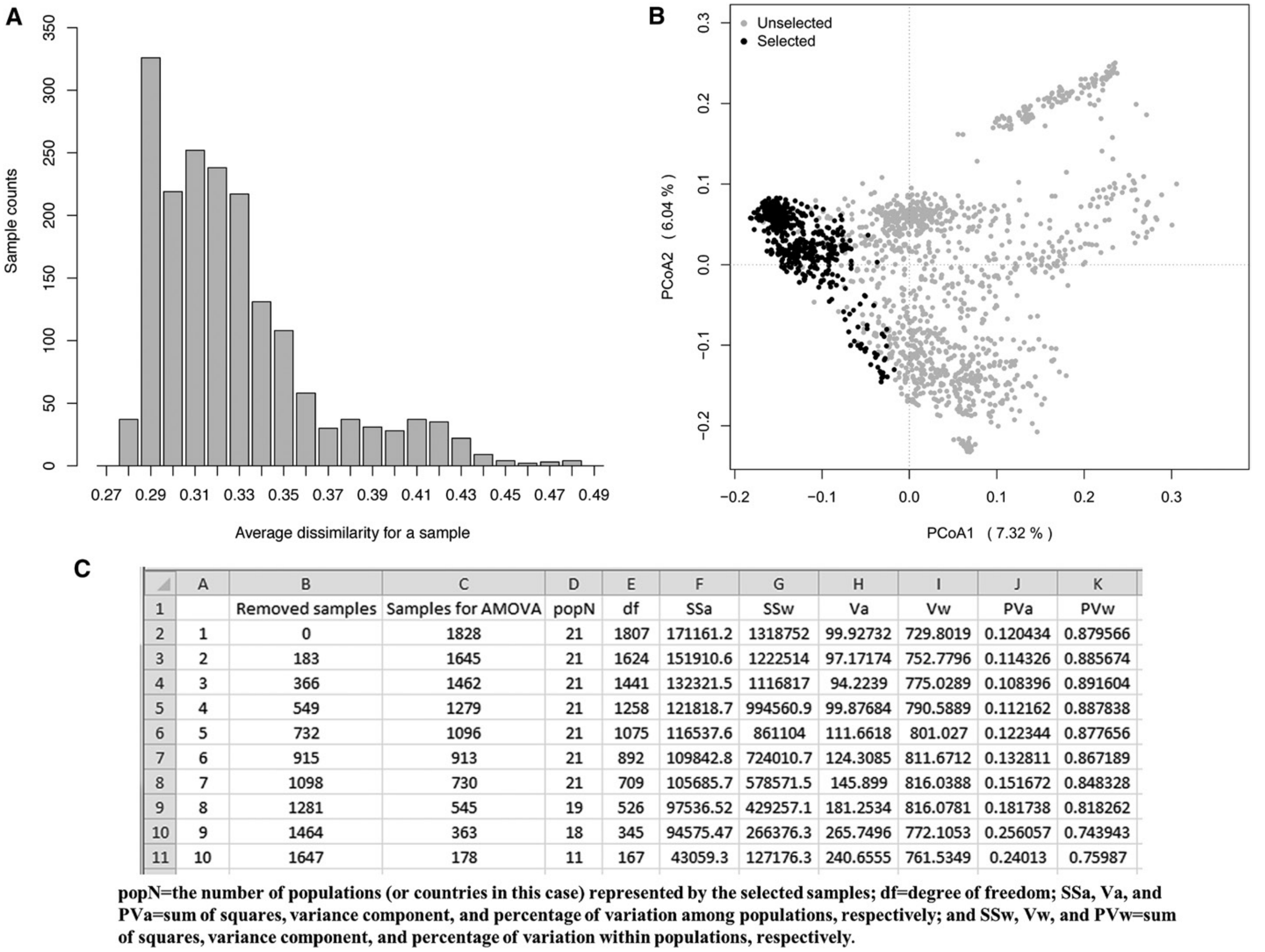
popN=the number of populations (or countries in this case) represented by the selected samples; df=degree of freedom; SSa, Va, and PVa=sum of squares, variance component, and percentage of variation among populations, respectively; and SSw, Vw, and PVw=sum of squares, variance component, and percentage of variation within populations, respectively.

Fig. 1. Illustration of AveDissR outputs for assessing genetic distinctness in an AFLP assay of 670 oat accessions. (A) Average dissimilarity (AD) frequency distribution in 670 oat accessions. (B) PCoA plot for the genetic associations of 461 accessions selected based on the highest AD values and original 670 accessions (selected accessions are shown in black). (C) Iterative AMOVA results from a stepwise selection of 10% in 670 oat samples; outputs show the changes in genetic differentiation (measured with PVa) with the increased selection of accessions with the highest AD values.

mean of 0.326 (Fig. 2A). Removing 549 accessions with the lowest AD values from the 1828 accessions did not show any large deviation from the original genetic differentiation presented in the assayed 1828 accessions, as shown with PVa values of 0.120434 in the first step and 0.112162 in the fourth step, respectively (Fig. 2C). The PCoA plot (Fig. 2B) showed the redundancy of the 549 samples with the lowest AD values in the original 1828 accessions, based on the revealed genetic associations. It is possible that one may need to rerun *AveDissR.r* with different parameters_setting files by adjusting *stepwise selection* values to identify a tentative group of genetically redundant accessions, as suggested by Fu (2006) with a nominal tolerance of 5% change in genetic differentiation from the originally assayed accessions. The tentative group of genetically redundant accessions could be used to verify those duplicated accessions previously identified (Song et al., 2015).

Application and limitation—AveDissR is an integrated statistical tool specifically developed for the AD approach to assist the assessment of genetic distinctness and/or redundancy in plant germplasm characterization. However, the AD approach per se is of limited resolution in defining genetically redundant or distinct groups, as clearly stated in Fu (2006) and shown in the analyses of oat and soybean data sets (see Figs. 1A, 2A). There is no definite criterion that one could develop and apply with AD values to identify genetically redundant or distinct groups of accessions. With additional information generated from iterative AMOVAs and PCoA plots, one could identify a tentative group of genetically distinct accessions for further evaluation and assessment in combination with passport, evaluation, and characterization data to develop a core subset from a large germplasm collection, as illustrated in the development of the flax core collection (Diederichsen et al., 2013). The tentative group of genetically distinct accessions per se may not necessarily be deemed as a core collection as defined by Brown (1989) for germplasm management and utilization. Similarly, the tentative redundant group can be used in combination with passport, evaluation, and characterization data to assist the identification of truly duplicated accessions in a germplasm collection for effective germplasm management and conservation. It is worth noting that the genetic redundancy acquired from an AveDissR application may not necessarily mean that the identified genetically redundant accessions are truly duplicated accessions.

Depending on research objectives, AveDissR can also be used to assess genetic distinctness or redundancy in experiments with selfing or outcrossing organisms in natural populations. For example, multiple individuals from one population (or accession in the sense of germplasm conservation) can be genotyped and analyzed using AveDissR, and genetic outliers (or the most genetically distinct individuals) from different populations can be identified. Similarly, the most genetically distinct populations over the geographic distribution of a species can be inferred for conservation and/or population genetic analyses. AveDissR can be applied to identify the genetically redundant (or similar) group of individuals from nearby natural populations and such group can allow for useful assessment of genetic relatedness in nearby populations. Thus, AveDissR



	A	B	C	D	E	F	G	H	I	J	K
1		Removed samples	Samples for AMOVA	popN	df	SSa	SSw	Va	Vw	PVa	PVw
2	1	0	1828	21	1807	171161.2	1318752	99.92732	729.8019	0.120434	0.879566
3	2	183	1645	21	1624	151910.6	1222514	97.17174	752.7796	0.114326	0.885674
4	3	366	1462	21	1441	132321.5	1116817	94.2239	775.0289	0.108396	0.891604
5	4	549	1279	21	1258	121818.7	994560.9	99.87684	790.5889	0.112162	0.887838
6	5	732	1096	21	1075	116537.6	861104	111.6618	801.027	0.122344	0.877656
7	6	915	913	21	892	109842.8	724010.7	124.3085	811.6712	0.132811	0.867189
8	7	1098	730	21	709	105685.7	578571.5	145.899	816.0388	0.151672	0.848328
9	8	1281	545	19	526	97536.52	429257.1	181.2534	816.0781	0.181738	0.818262
10	9	1464	363	18	345	94575.47	266376.3	265.7496	772.1053	0.256057	0.743943
11	10	1647	178	11	167	43059.3	127176.3	240.6555	761.5349	0.24013	0.75987

popN=the number of populations (or countries in this case) represented by the selected samples; df=degree of freedom; SSa, Va, and PVa=sum of squares, variance component, and percentage of variation among populations, respectively; and SSw, Vw, and PVw=sum of squares, variance component, and percentage of variation within populations, respectively.

Fig. 2. Illustration of AveDissR outputs for assessing genetic redundancy in an assay of 1828 soybean accessions. (A) Average dissimilarity (AD) frequency distribution in 1828 soybean accessions. (B) PCoA plot for the genetic associations of 549 accessions selected based on the lowest AD values and original 1828 accessions (selected accessions are shown in black). (C) Iterative AMOVA results from a stepwise selection of 10% in 1828 soybean samples; outputs show the changes in genetic differentiation (measured with PVa) with the increased removal of accessions with the lowest AD values.

has the potential to be applied in some genetic studies of natural populations, particularly for conservation biology. However, its application has not reached to genetic studies of natural populations yet and its usefulness in these studies needs further exploration.

CONCLUSIONS

The developed AveDissR function integrates three components of average dissimilarity calculation, iterative AMOVA, and PCoA as a package to assist the assessment of genetic distinctness or genetic redundancy in molecular characterization of plant germplasm. It can be applied to analyze five different types of dominant or codominant marker data, generate a useful set of output files for germplasm assessment, and run in R environments under different computer platforms. However, the R function, even with parallel computing capability, could still take hours or even days to analyze large marker data sets, particularly with missing values, as pairwise dissimilarity calculation for a large number of germplasm accessions (e.g., 10,000 accessions × 20,000 SNPs) is computationally intensive. Also, the average dissimilarity approach per se has its weaknesses and may not allow for a full identification of genetically redundant accessions (Fu, 2006). Nevertheless, the outputs from AveDissR should be useful for the assessment of genetic distinctness and/or genetic redundancy in plant germplasm characterization.

LITERATURE CITED

Brown, A. H. D. 1989. Core collections: A practical approach to genetic resources management. *Genome* 31: 818–824.

Chavarriaga-Aguirre, P., M. M. Maya, J. Tohme, M. C. Duque, C. Iglesias, M. W. Bonierbale, S. Kresovich, and G. Kochert. 1999. Using microsatellites, isozymes and AFLPs to evaluate genetic diversity and redundancy in the cassava core collection and to assess the usefulness of DNA-based markers to maintain germplasm collections. *Molecular Breeding* 5: 263–273.

Dean, R. E., J. A. Dahlberg, M. S. Hopkins, S. E. Mitchell, and S. Kresovich. 1999. Genetic redundancy and diversity among ‘Orange’ accessions in the US national sorghum collection as assessed with simple sequence repeat (SSR) markers. *Crop Science* 39: 1215–1221.

Diederichsen, A., P. M. Kusters, D. Kessler, Z. Bainas, and R. K. Gugel. 2013. Assembling a core collection from the flax world collection maintained by Plant Gene Resources of Canada. *Genetic Resources and Crop Evolution* 60: 1479–1485.

Engels, J. M. M. 2004. Plant genetic resources management and conservation strategies: Problems and progress. *Acta Horticulturae* 634: 113–125.

Engels, J. M. M., and L. Visser. 2003. A guide to effective management of germplasm collections. IPGRI Handbooks for Genebanks No. 6. International Plant Genetic Resources Institute, Rome, Italy.

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131: 479–491.

FAO. 2010. The second report on the state of the world's plant genetic resources, p. 370. Food and Agriculture Organization of the United Nations, Rome, Italy.

Frankel, O. H. 1984. Genetic perspectives of germplasm conservation. *In* W. K. Arber, K. Llimensee, W. J. Peacock, and P. Starlinger [eds.], Genetic manipulation: Impact on man and society, 161–170. Cambridge University Press, Cambridge, United Kingdom.

Fu, Y. B. 2006. Redundancy and distinctness in flax germplasm as revealed by RAPD dissimilarity. *Plant Genetic Resources* 4: 117–124.

Fu, Y. B., and G. W. Peterson. 2012. Developing genomic resources in two *Linum* species via 454 pyrosequencing and genomic reduction. *Molecular Ecology Resources* 12: 492–500.

Fu, Y. B., G. W. Peterson, D. Williams, K. W. Richards, and J. Mitchell Fetch. 2005. Patterns of AFLP variation in a core subset of cultivated hexaploid oat germplasm. *Theoretical and Applied Genetics* 111: 530–539.

Hamming, R. W. 1950. Error detecting and error correcting codes. *Bell System Technical Journal* 29: 147–160.

Karp, A. 2002. The new genetic era: Will it help us in managing genetic diversity? *In* J. M. M. Engels, V. R. Rao, A. H. D. Brown, and M. T. Jackson [eds.], Managing plant genetic diversity, 43–56. International Plant Genetic Resources Institute, Rome, Italy.

Kisha, T. J., and C. S. Cramer. 2011. Determining redundancy of short-day onion accessions in a germplasm collection using microsatellite and targeted region amplified polymorphic markers. *Journal of the American Society for Horticultural Science* 136: 129–134.

Motilal, L. A., D. Zhang, S. Mischke, L. W. Meinhardt, and P. Umaharan. 2013. Microsatellite-aided detection of genetic redundancy improves management of the International Cocoa Genebank, Trinidad. *Tree Genetics & Genomes* 9: 1395–1411.

Peakall, R., and P. E. Smouse. 2006. GenAlEx 6: Genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6: 288–295.

Peakall, R., and P. E. Smouse. 2012. GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research–An update. *Bioinformatics (Oxford, England)* 28: 2537–2539.

Peterson, G. W., Y. Dong, C. Horbach, and Y. B. Fu. 2014. Genotyping-by-sequencing for plant genetic diversity analysis: A lab guide for SNP genotyping. *Diversity (Basel)* 6: 665–680.

Phippen, W. B., S. Kresovich, F. G. Candelas, and J. R. McFerson. 1997. Molecular characterization can quantify and partition variation among genebank holdings: A case study with phenotypically similar accessions of *Brassica oleracea* var. *capitata* L. (cabbage) 'Golden Acre.' *Theoretical and Applied Genetics* 94: 227–234.

R Core Team. 2016. R: A language and environment for statistical computing. Version 3.3.1 (2016-06-21). R Foundation for Statistical Computing, Vienna, Austria. Website https://www.R-project.org/ [accessed 23 July 2016].

Rohlf, F. J. 1997. NTSYS-pc 2.1. Numerical Taxonomy and Multivariate Analysis System. Exeter Software, Setauket, New York, USA.

SAS Institute. 2004. The SAS System for Windows V8.02. SAS Institute, Cary, North Carolina, USA.

Sokal, R. R., and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38: 1409–1438.

Song, Q., D. L. Hyten, G. Jia, C. V. Quigley, E. W. Fickus, R. L. Nelson, and P. B. Cregan. 2015. Fingerprinting soybean germplasm and its utility in genomic research. *G3: Genes, Genomes, Genetics* 5: 1999–2006.

van Hintum, T. J. L., A. H. D. Brown, C. Spillane, and T. Hodgkin. 2000. Core collections of plant genetic resources. IPGRI Technical Bulletin No. 3. International Plant Genetic Resources Institute, Rome, Italy.

Virk, P. S., H. J. Newbury, M. T. Jackson, and B. V. Ford-Lloyd. 1995. The identification of duplicate accessions with a rice germplasm collection using RAPD analysis. *Theoretical and Applied Genetics* 90: 1049–1055.

Wang, C., W.-H. Kao, and C. K. Hsiao. 2015. Using Hamming Distance as information for SNP-sets clustering and testing in disease association studies. *PLoS ONE* 10: e0135918.

Waycott, W., and S. B. Fort. 1994. Differentiation of nearly identical germplasm accessions by a combination of molecular and morphological analyses. *Genome* 37: 577–583.

Yang, M.-H., and Y. B. Fu. 2017. Supplementary data for: AveDissR: An R function for assessing genetic distinctness and genetic redundancy. Figshare, http://dx.doi.org/10.6084/m9.figshare.5082451.