Availability of computer code and algorithm

The process for statistical analysis using not published software, but codes inside the free R software environment, consisted in five stages, each one conducted separately for the three biological categories (CWR species, cultivated tetraploid and cultivated hexaploid). All of them were conducted using appropriate packages from the free software R, the five stages were:

1. INPUT. Read the csv files sent by the laboratory, usually more than three hundred of files. Calculate allele frequencies and proportion of missing values for the SNP files, and only translate to R the Presence Absence Values (PAV).
2. Distances. Calculate the Modified Rogers genetic distance or the Jaccard distance
3. Multi-Dimensional Scaling for graphical representation of distances
4. Analysis of Molecular Variance (AMOVA) for SNP’s data.
5. Building a Core subset using the D-Method: classification (clustering), selecting the appropriate number of groups proportional to the diversity per group, building one thousand of “candidate Cores” and selecting the most diverse of them to be the “core subset”.

Code for stages 1, 3, 4 are usual R scripts (read/transform or create new variables/write), and even being necessary stages are not the central scripts on which our results are based, stages 2 (particularly the Modified Roger Distance, *mrd*) and 5 (applying the D-method for building Core Subsets) which imply a classification are attached to this note. Besides we included real data if reviewers wish test the scripts. Upon request we are available to send the scripts associated with all the process

STAGE 2.

For stage 2 we used a “small” dataset, frF1F2.RData, containing allele frequencies from 4403 Wild Relative accessions and 112562 alleles, we used that data to apply the scripts calculating the *mrd* genetic distance. There are three scripts to be run in the presented sequence:

01-SNPparallelMRDdistV7.R: starts reading the frequencies file and ends writing15 R-binary files (vectors) mrd2Vec\_1\_1.RData to mrd2Vec\_4\_5.RData containing squared *mrd* distances between all pair of accessions grouped in five sets (four of 1K and one of 403 accessions, respectively) to split and make possible the calculus in a parallel process. Results are in the armrd2Vec.tar file.

02-mrdVec.R: joins the 15 mentioned files and create a single vector containing the 9,691,003 distances between all pair of accessions (spuared root of mrd^2 mentioned distances). The process is a parallel and quick one. The result is in the R-binary mrdVec.RData file.

03-indexYvtomtodNew-WR.R: starting from mrdVec it creates an index to group distances and two matrices containing distances: mrdMatNew.RData (symmetric matrix of distances) and mrdDistnew.RData (lower triangular matrix of distances).

The calculus of distances is a long process due to the number of operations, that is the reason we used the small dataset.

STAGE 5. Classification and core subset selection

For this stage we used a bigger distances file (54636 accessions extracted from the original hexaploid data accessions) containing 1,492,518,930 distances. Four scripts are showed:

01-snpCluster-hpc.R: starts using the mrd distances matrix, uses the fastCluster R package, do the clustering, generates the classification of accessions into 2,3,…,50 groups (mrdgr2-50.csv file) to be used in the evaluation of each group characteristics (using the fpc R package) to select an appropriate number of groups. This stage is not a parallel process, the statistics to evaluate the different groupings (2 to 50 groups) are average of distances between and within groups, within cluster sum of squares and the pseudo-F statistic. Result were pasted to excel file to do the graphical and comparative analysis.

02-describeGroups.R: using the mentioned distances matrix and the selected grouping (classification into 7 groups in this case) creates the groupsDesc.csv file containing the size of each group, its mrd average distance and the standard deviation of distances per group. That information is used to assign the sample size to be sampled from each group using a stratified random sampling process. The assignation is done proportional the *mrd* average value per group.

03-simucores.R: Applies the D-method (reference in the paper). Uses the distances matrix, the groups (7 in this case) and the sample size assignation (vector *ni* in the script), obtains 1000 stratified random samples (candidate subsets) and their *mrd* average distances, select the best candidate (showing the maximum of average distances) write the core selection and prints the core and population description to be compared.

04-EvalCorePop.R: Using the allele frequencies original file and the selected core calculates the diversity indices for the core subset and the population. These indices will be used ahead to the AMOVA and other analyses.