short manual

DuplicateFinder

DuplicateFinder is a tool to support you in your search for duplicates and most appropriate accessions, it doesn't do the job for you. It will still be difficult and time consuming to find the duplicates. We just hope DuplicateFinder will help you in this effort and save you a little time.

Introduction

DuplicateFinder is a little piece of software to help identify potential duplicate PGR samples on the basis of passport data. It is simple to use, and will help anyone analysing a Central Crop Database, or a local PGR documentation system by creating groups of accessions that are likely to contain the potential duplicates. In the end the user will have to decide about which accessions (s)he thinks are actual duplicates.

The software comes in a MS Excel environment: it is a spreadsheet with macro's. The user only has to copy the passport data in the spreadsheet, and run the macro's either (1) to identify the potential duplicates of a selected accession or (2) to create groups of material which can contain the duplicates.

Restrictions and limitations

1 – The passport data have to be formatted according to the Multi Crop Passport Descriptor (MCPD) list. Or, actually, the software currently only uses the descriptors ACCENUMB, ACCENAME, COLLNUMB, DONORNUMB and OTHERNUMB, so these fields have to have the right headers in the sheet so that the software can recognise them.

2 – The fact that only these five fields are used, implies that other fields which might contain clues about duplication, such as taxonomic and origin location fields, are not used. Duplicates that can be identified only on these fields will not be identified by the software!

3 – The first time the software is used, it will take some time (depending on the hardware and size of the dataset, a few minutes max) to prepare the data structures used for the identification.

4 - The creation of potential duplication groups can take much longer (up to a few hours for large data sets), and can only be run on data sets with 15000 accessions or less. If the data set is larger, the user has to restrict the number of records to be processed. When large numbers of records are to be processed it can be useful to clear the memory buffers by saving, closing and re-opening the file.

5 – The software can change the format of your data sheet (layout, colour, size, etc.). It will never change the data values, except in the macros under the 'Adjust content' option!

Steps

1 – Open DuplicateFinder

Open the spreadsheet DuplicateFinder.xlsm in such a way that macro's can be run. This might require some changes in the security setting of Excel, but most likely only involves clicking 'accept' at some stage. (This manual is assuming you are using Excel 2010 or later, if you are using an older version of Excel, you might have to look for the macros mentioned below since the menu structure might be slightly different – but everything will function also in earlier versions.)

2 – Copy your data

Make sure the data you want to analyse are in a spreadsheet, ready to be copied in the sheet 'DATA' of DuplicateFinder. Delete the sample data in the sheet 'DATA' and copy you're your data in.

3 – Check format of your data

If you are not sure about the formatting, check the sheet 'MCPD List' of DuplicateFinder. It lists the descriptors of the Multi Crop Passport Descriptor (MCPD) list. Make sure your data in the sheet 'DATA' have at least proper headings for the descriptors ACCENUMB, ACCENAME, COLLNUMB, DONORNUMB and OTHERNUMB. If you like you can have DuplicateFinder check the format by running the macro 'ValidateAllColumns' (click menu option 'Add-Ins', click option 'DuplicateFinder' in the Add-Ins ribbon, chose 'Validate columns' and 'Validate all columns'). In the sheet 'Report' you will see what columns were found or missing, what values were missing, wrongly formatted or wrongly coded. In the data sheet the recognized columns will have bold headers, and the wrong values in these columns will be given a red background. The macros provided under the option 'Adjust Content' will help you reformatting. If you have corrected headers or fields and want to check again, you can restrict the check to the column you changed (run the macro 'ValidateOneColumn' by choosing the option 'Validate one column'). If you want to change the formatting of some of the MCPD columns, check the menu option 'Adjust content', but be aware that these macros do change the content of the data !

4 – Find potential duplicates of one accession

Select a cell in the record of the accession you want to match with the others, and run the macro 'FindDuplicatesOneAcc' (click menu option 'Add-Ins', click option 'DuplicateFinder' in the Add-Ins ribbon, chose 'Find duplicates' and 'Find duplicates for one accession'). For the first search the software might need some time to prepare the required data structures (few minutes max). The software will create two new columns DFSim and DFIDno, if they were not created before. DFIDno will contain temporary unique ID numbers of each accession – you can ignore it, it will be hidden after the calculations have been finished. The column DFSim will show the similarity between the selected accession (at the top of the sheet) and others who are displayed in decreasing order of similarity.

5 – Create potential duplication groups

If you run the macro ‘MakeDuplicateGroups’ (click menu option ‘Add-Ins’, click option ‘DuplicateFinder’ in the Add-Ins ribbon, chose ‘Find duplicates’ and ‘Make duplicate groups’), DuplicateFinder will create groups with similar accessions. Since each accession needs to be compared with each other accession, this might take a while (up to a few hours, depending on hardware and number of records). The result of all these calculations will be a new column called DFGrp, where similar accessions will be given the same group number. Accessions that were not clustered with others will have no group number.

Careful

DuplicateFinder is designed to identify potential duplicates, not to provide an environment to edit data. However, if you do decide to change the data, and want to continue searching duplicates or creating groups, you need to recalculate the similarities between the accessions. For this purpose you should run the macro ‘RecalculateData’ (click menu option ‘Add-Ins’, click option ‘DuplicateFinder’ in the Add-Ins ribbon, chose ‘Find duplicates’ and ‘Recalculate data’). Be aware that the grouping you calculated before is not changed, unless you recalculate it.

Tips for use

- First play a few minutes with the software, using the 1000 sample records that come with the spreadsheet, and see the possibilities. Be aware that everything takes longer if there are more records – the time to create groups is roughly quadratic to the number of accessions – twice as many accessions takes four times as long.

- Create a column GROUP (or something similar) to store the groups you accepted or identified yourself, or a column STATUS, to indicate if an accession is a ‘Most Appropriate’ or a ‘Probable Duplicate’. Based on the results of running the macro’s you can fill and change the values in these columns.

- If the number of accessions in your data set exceeds 15000, you should select a homogeneous group (one taxon, only cultivars, etc.) to run the ‘MakeDuplicateGroups’ macro since it cannot handle more accessions.

- If you want to run the ‘MakeDuplicateGroups’ macro, and the number of accessions is high, consider starting it before going home or to a meeting. Make sure that the ‘Power savings options’ of your computer doesn’t prevent it from continuing to work when you leave the room.

Acknowledgements

DuplicateFinder was developed in the framework of, and with financial support of the AEGIS initiative of the European Cooperative Programme for Plant Genetic Resources (ECPGR) by the Centre for Genetic Resources, The Netherlands (CGN - Roel Hoekstra, Theo van Hintum, Frank Menting and other staff) with support of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK - Helmut Knüpffer) and Julius Kühn Institute (JKI - Christoph Grehmeier), both in Germany.

FEEDBACK

If you have problems using the DF, or find bugs that we didn't find, please contact the developers: Roel Hoekstra (roel.hoekstr@wur.nl) or Theo van Hintum (theo.vanhintum@wur.nl).