

# BRB-ArrayTools: DAPfinder Plug-in

## 1. Introduction

The *DAPfinder* plug-in for BRB-ArrayTools helps microarray researchers identify statistically significant differences in gene-gene association between two classes of microarray data. These significant differences in gene-gene association can be referred to as “Differentially Associated Pairs” or DAPs. The plug-in uses a Fisher’s Z test to compare gene-gene Pearson, Spearman or Kendall correlations between two classes and it provides an option to compare gene-gene mutual information, Pearson, Spearman or Kendall correlations between two classes with a permutation test. Several filtering options are available to help researchers identify the most relevant genes, including outlier removal options and coefficient of variation (CV) filtering. Results are stored in a MS Excel file or tab-delimited text file, while user settings are printed in an HTML report. This plug-in was created by Jeff Skinner, M.S., Sudhir Varma, Ph.D., Vivek Gopalan, Ph.D. and Yenram Huyen, Ph.D. of the Bioinformatics and Computational Biosciences Branch (BCBB) at the National Institute of Allergy and Infectious Disease (NIAID) in collaboration with Andrey Morgun, M.D. Ph.D. of the Laboratory of Cellular and Molecular Immunology (LMCI) at NIAID, Yuri Kotliarov, Ph.D. of the Neuro-Oncology Branch (NOB) at the National Cancer Institute (NCI) and Anatoli Yambartsev, Ph.D. of the Instituto de Matemática e Estatística (IME) at Universidade de São Paulo (USP). The BRB-ArrayTools package was developed by Dr. Richard Simon and Amy Peng Lam of the Biometric Research Branch (BRB) at NCI. This manual describes how users can install and use the *DAPfinder* BRB-ArrayTools plug-in.

## 2. Install the BRB-ArrayTools package

First, users need to install the BRB-ArrayTools package. Visit the NCI BRB-ArrayTools website (<http://linus.nci.nih.gov/BRB-ArrayTools.html>) and follow their download instructions. Please note the BRB-ArrayTools package requires MS Excel, MS Visual Basic for Applications (VBA), the R statistical computing package, Java and other commonly used PC applications. The *DAPfinder* plug-in was created using R scripts that require R version 2.11.0 or higher and a few unusual R packages, but the plug-in can automatically update your R installation and install the necessary packages if your computer is connected to the internet.

## 3. Obtain the DAPfinder plug-in

If you are reading this manual, then you should already have a copy of the *DAPfinder* plug-in. If you do not have the *DAPfinder* plug-in, then please contact [ScienceApps@niaid.nih.gov](mailto:ScienceApps@niaid.nih.gov) to obtain a copy. The plug-in is distributed as a compressed zip file titled *DAPfinder.zip*. The unzipped folder contains 8 files: *DAPfinder.plug*, *DAPfinder.R*, *DAPfinder.txt*, *dialogs.R*, *functions.R*, *perm.all.R*, *permute\_gene.dll* and *DAPfinder Manual.pdf*.

#### 4. Install the DAPfinder plug-in

Open the Windows file directory `c:\Program Files\ArrayTools\plugins` and paste the unzipped *DAPfinder* folder into the *plugins* folder of the directory specified above (Figure 1). Please note that your directory may be slightly different.

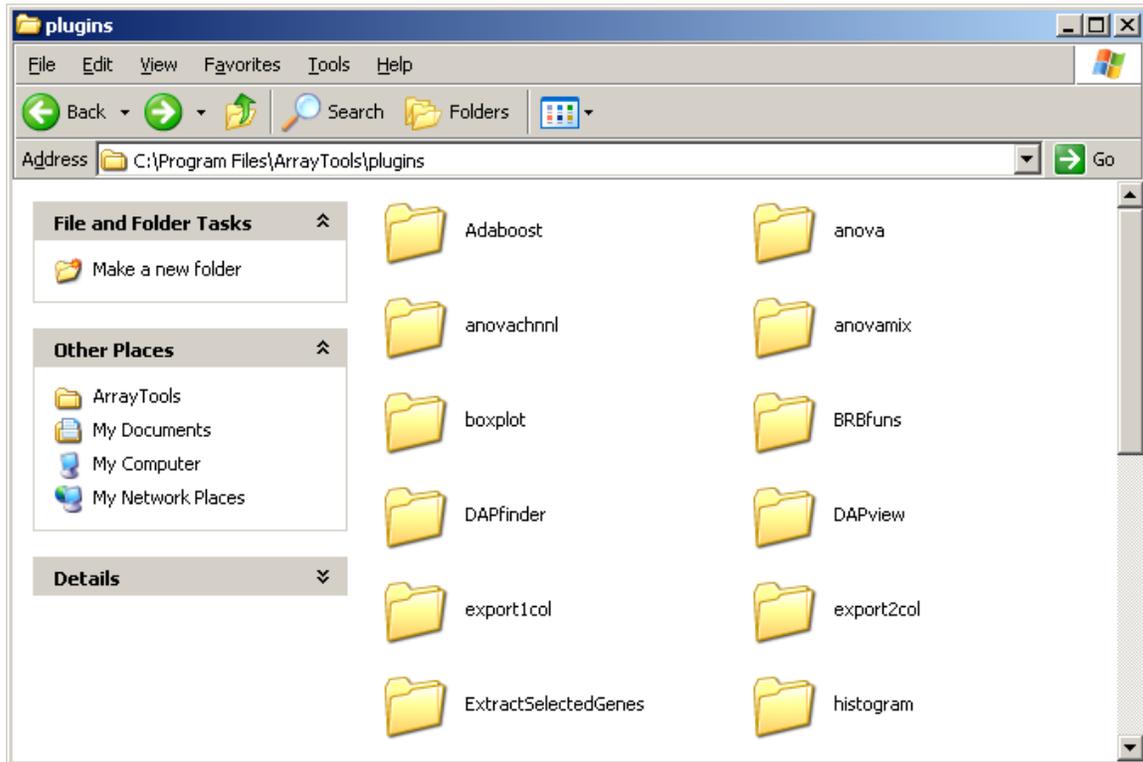


Figure 1. Paste the *DAPfinder* folder in the specified directory.

Verify that the *DAPfinder* folder has been correctly pasted into the ArrayTools plugins folder and make sure its contents are intact. Open MS Excel to load the plug-in into BRB-ArrayTools. Click > **ArrayTools** > **Plugins** > **Load Plug In** to install the *DAPfinder* plug-in from the *Load Plug In* menu of BRB-ArrayTools (Figure 2).

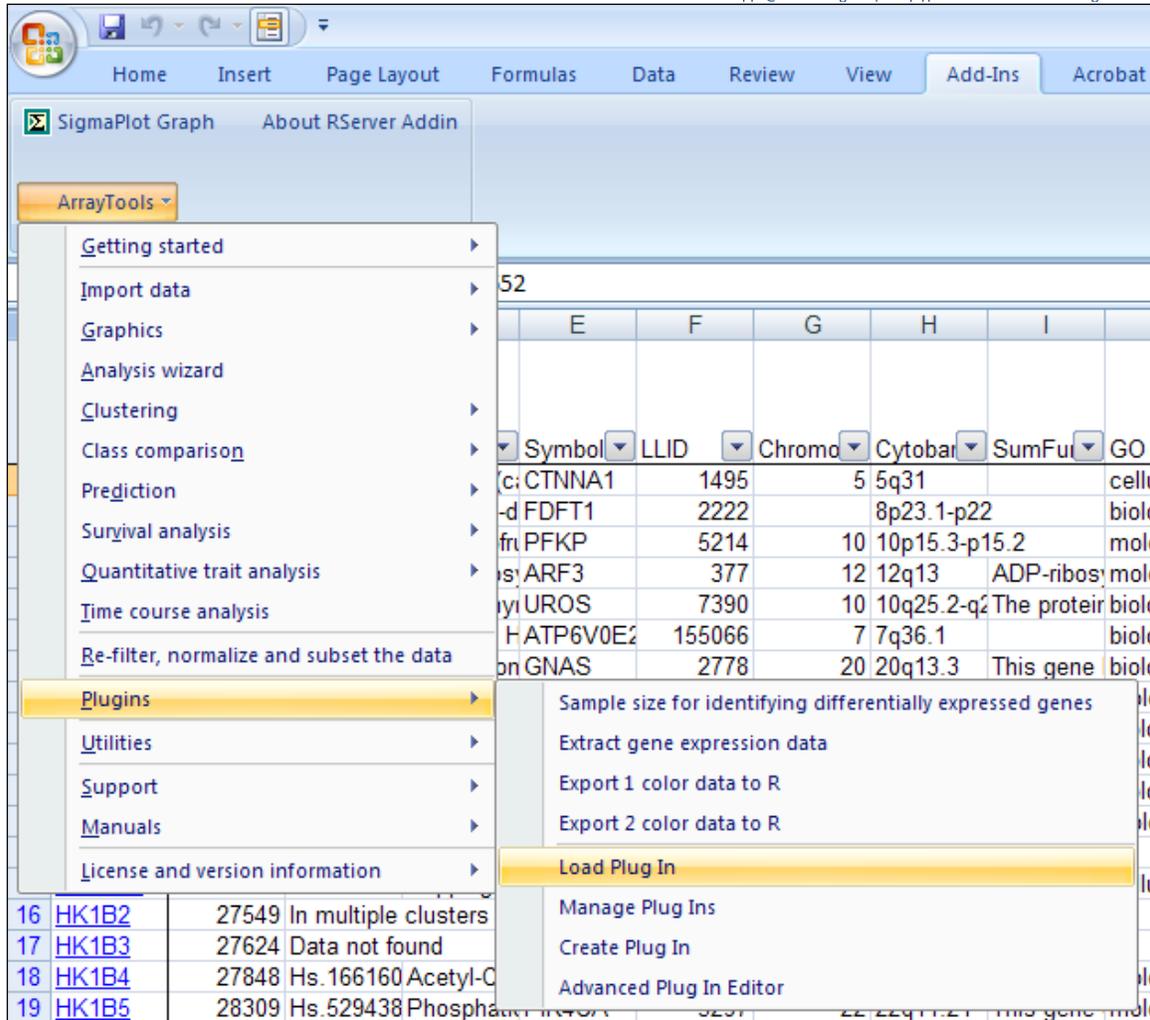


Figure 2. Click > **ArrayTools** > **Plugins** > **Load Plug In** to load the plug-in.

Click the Browse button to locate the *DAPfinder.plugin* file in the directory:

```
c:\Program Files\ArrayTools\plugins\DAPfinder\DAPfinder.plugin
```

Check the box to add your plug-in to the menu and click “OK” to finish (Figure 3). The plug-in should open automatically and it will be located in the > **ArrayTools** > **Plugins** > **DAPfinder** menu for future use (Figure 4).

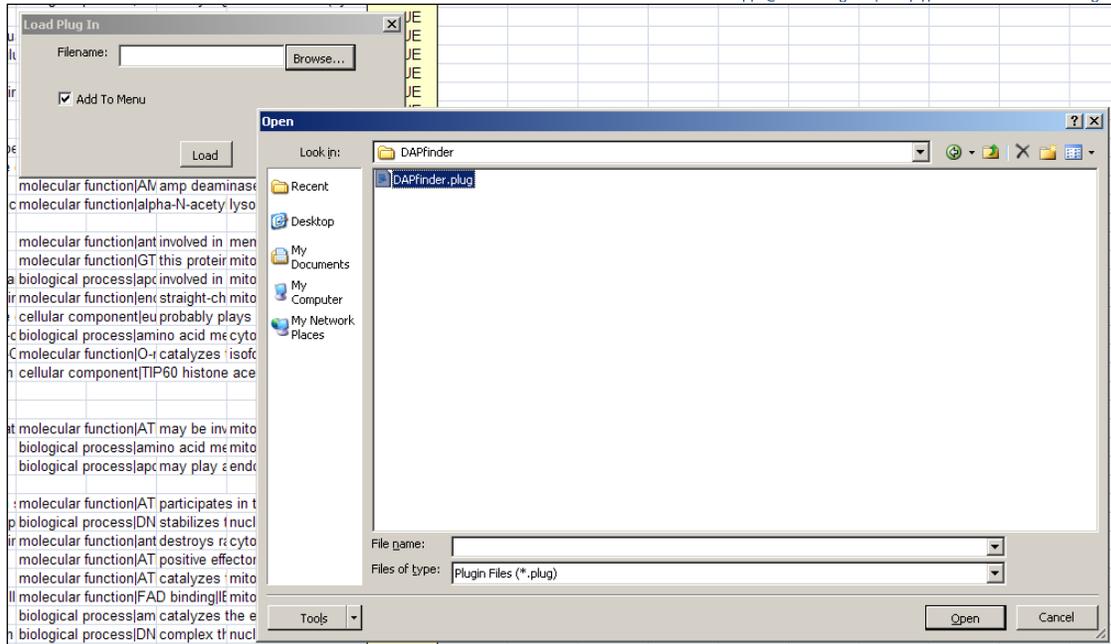


Figure 3. Locate the *DAPfinder.plugin* file.

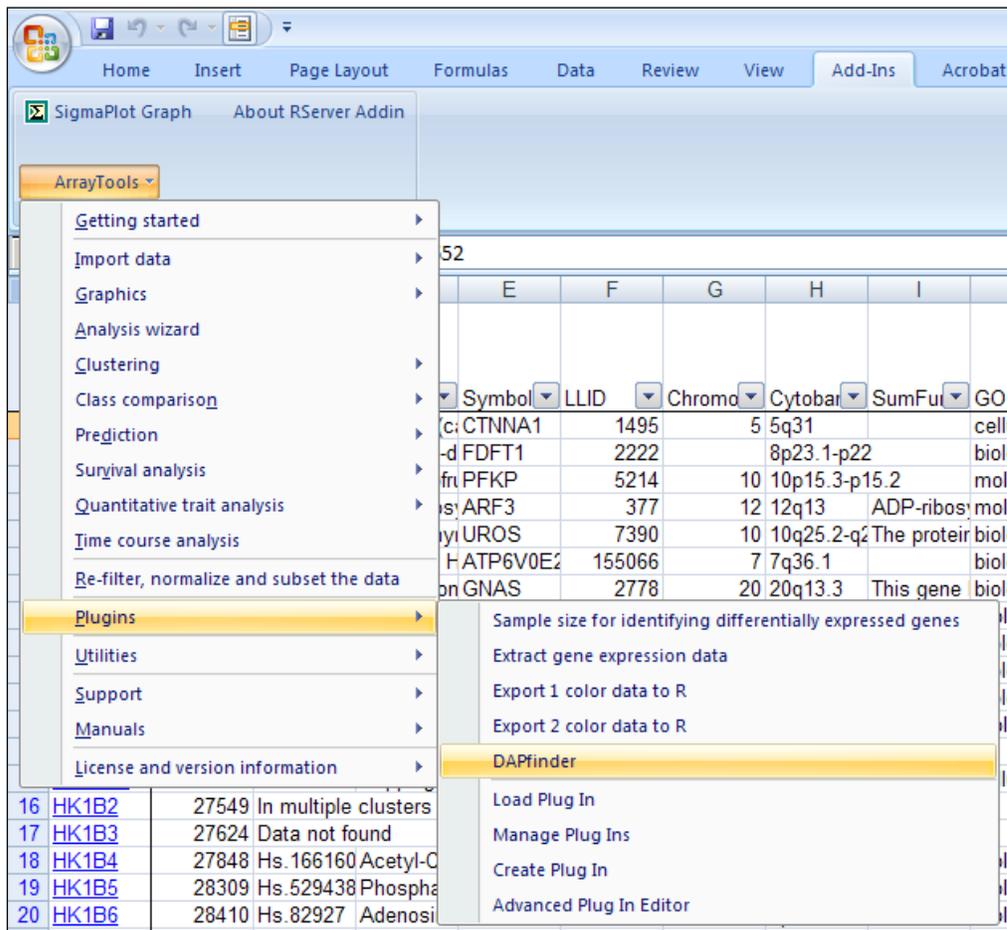


Figure 4. The > **ArrayTools** > **Plugins** > **DAPfinder** menu.

5. Use the DAPfinder plug-in.

Follow the instructions from the BRB-ArrayTools documentation to open a BRB-ArrayTools data set. Each BRB-ArrayTools data set will be a MS Excel workbook with at least three worksheets (Figure 5). One sheet will be the *Experiment descriptors* sheet, which stores information about the treatments and experimental designs applied to each microarray chip. Another sheet is the *Gene identifiers* sheet, which stores information about the individual genes stored on each microarray chip. Often, the Gene identifiers sheet will contain a column of unique gene ID's (i.e. *UniqueID*) or GenBank accession numbers (i.e. *GB acc*). The third required sheet is the *Filtered log ratio* or *Filtered log intensity* sheet, which stores the actual gene expression values for each gene (i.e. each row) and each microarray chip (i.e. each column). There may be an optional *Gene annotations* sheet or other user specified sheets.

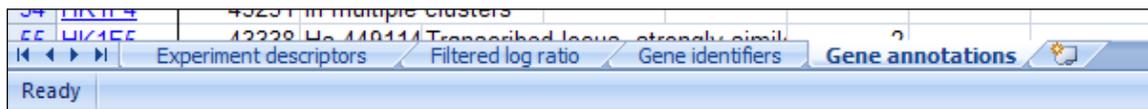


Figure 5. Four worksheets from a BRB-ArrayTools dataset.

There are many useful functions among the stock BRB-ArrayTools tools, but users of the *DAPfinder* plug-in may be most interested in the *Re-filter, normalize and subset the data* option (Figure 6). Click > **ArrayTools** > **Re-filter, normalize and subset the data**, then select appropriate options to apply normalization schemes; to filter the gene expression values based on spot filters, fold change, variation or percent missing values; or to select known gene sets for analysis. These BRB-ArrayTools filter settings can be accessed in the *DAPfinder* plug-in.

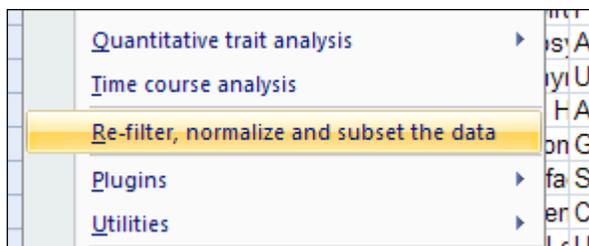


Figure 6. The BRB-ArrayTools *Filter and subset the data* option.

Click > **ArrayTools** > **Plugins** > **DAPfinder** to open the *DAPfinder* window in BRB-ArrayTools (Figure 7). The menu should contain 12 fields to specify user inputs from BRB-ArrayTools. The fields are typically drop-down menus or text boxes.

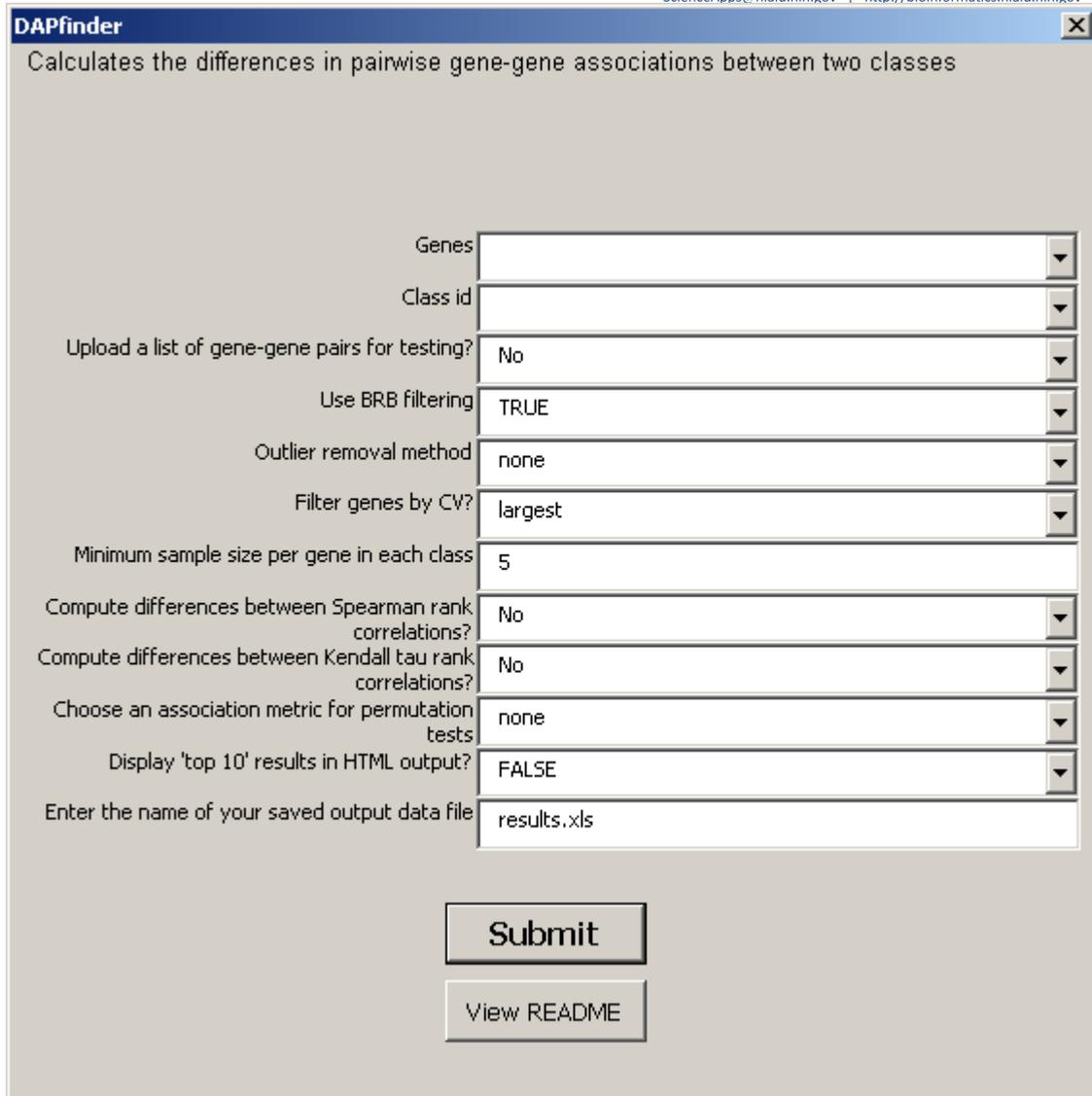


Figure 7. The *DAPfinder* window in BRB-ArrayTools.

Note that you can hover your cursor over some fields to reveal helpful hints about what the fields mean or what inputs may be appropriate (Figure 8). E.g. The hint for the *Upload a list of gene-gene pairs for testing?* field warns users that the uploaded list must be a tab-delimited text file.



Figure 8. Find user hints by hovering your cursor over fields.

Click the *View README* button at the bottom of the *DAPfinder* window to view the README text file (Figure 9). The README text file contains important information about the plug-in, including definitions of all 12 fields of the *DAPfinder* window. The README file may be a valuable resource if this manual is not available. Enter appropriate values into all 12 fields and click the *Submit* button to complete the analysis and view your results.

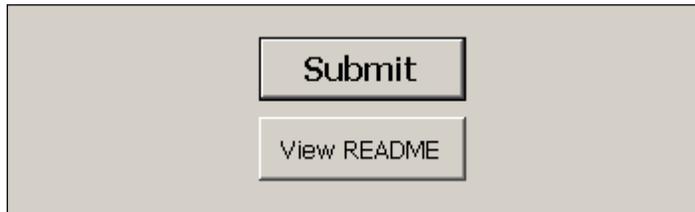


Figure 9. The *View README* and *Submit* buttons.

## 6. Specifying Fields in the *DAPfinder* Main Window.

The *DAPfinder* main window is the user interface created by the *DAPfinder.plug* file, and it allows you to specify the most basic parameters for your analyses. The first field on the *DAPfinder* window is the *Genes* field, which is used to specify one column of unique gene names from the *Gene identifiers* sheet of a BRB-ArrayTools dataset (Figure 10). Please note that two rows cannot share the same gene name. If your data set contains two rows that share the same name, it should generate a fatal error. Also note that all the information from the *Gene identifiers* sheet will be included in the final output, but only one column of gene names may be selected for the plug-in analyses.



Figure 10. The *Genes* field for one column of microarray gene names.

The second field is the *Class id* field, which is used to specify one column of class identifiers from the *Experiment descriptors* sheet of a dataset (Figure 11). Class ids are used to separate the microarray chips into two different classes, so we can compare the gene-gene associations between these two classes. The two classes could denote any two different groups (e.g. WT vs. KO mice, treated vs. placebo patients, male vs. female, etc). Indicator variables (e.g. 0 vs. 1) or text strings (e.g. WT vs. KO) may be used, but there may only be two classes. Variables with 3 or more classes will produce an error.



Figure 11. The *Class id* field for one column of class identifiers.

The *Upload a list of gene-gene pairs for testing?* field allows users to upload a list of specific gene-gene pairs to be tested by the plug-in (Figure 12). This list of gene-gene pairs must be a tab-delimited text file with exactly two columns representing the source and target genes of each gene-gene pair, respectively (Figure 13). It should be easy to create such a list from MS Excel using > **File** > **Save As** with *Save as type* set to *Text (Tab delimited)*. You could also copy and paste two columns representing your gene-gene pairs directly into MS Notepad. If you choose to upload a file of specific gene-gene pairs, the plug-in will only display test output for the gene-gene pairs you have specified. However, keep in mind that other filtering settings can be applied to this gene-gene list, so carefully choose your remaining filter settings.



Figure 12. The *Upload a list of gene-gene pairs for testing?* field for association tests of user-specified gene-gene pairs.

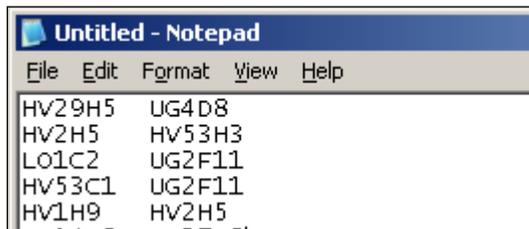


Figure 13. A tab-delimited text file of gene-gene pairs

The *Use BRB filtering* field allows users to choose whether they want to use the filter settings from the BRB-ArrayTools *Filter and subset the data* option (Figure 14). This field is a drop-down menu, where the choice *TRUE* indicates that you want to use the BRB filter settings and the choice *FALSE* indicates that you do not want to use the BRB filter settings. These filtering options allow you to apply normalization and fold change filtering or choose gene subsets, as described above.



Figure 14. The *Use BRB filtering* field to use *Filter and subset the data* settings.

The *Outlier removal method* field allows users to choose one of five univariate outlier removal methods (Figure 15). The *IQR – pooled* choice pools both classes of data into one group and removes suspected outlier gene expression values outside the range  $[Q1 - k \cdot IQR, Q3 + k \cdot IQR]$ , where  $Q1$  is the 25<sup>th</sup> percentile,  $Q3$  is the 75<sup>th</sup> percentile,  $IQR$  is the interquartile range and  $k$  is a constant (default  $k = 1.5$ ). The *IQR – unpooled*

choice separates the data into two classes based on the *Class id* field, then applies the same IQR scheme described above. The *SD – pooled* choice pools both classes of data into one group and removes suspected outlier gene expression values outside the range  $[\text{mean} - k \cdot \text{SD}, \text{mean} + k \cdot \text{SD}]$ , where mean is the sample mean, SD is the sample standard deviation and  $k$  is a constant (default  $k = 4$ ). The *SD – unpooled* choice separates the data into two classes based on the *Class id* field, then applies the same SD scheme described above. If users choose *none*, then no outlier removal is performed. Please note that selecting IQR or SD outlier removal methods will generate additional pop-up window prompts after you click “Submit”



Figure 15. The *Outlier removal method* field.

The *Filter genes by CV?* field allows users to select genes with the largest or smallest coefficient of variation from each class (Figure 16). The coefficient of variation (CV) is the standard deviation of expression values for each gene divided by the mean expression value for that gene. Genes with large CV values will have a lot of variation in expression levels, but they may be noisy, while genes with the smallest CV values will have little to no variation in expression values. The field has a drop-down menu with choices for *largest CV*, *smallest CV* or *none*. The default setting is *largest CV* filtering. The choice *none* indicates that no CV filtering is used. If the user chooses largest or smallest CV filtering, then the user will need to respond to a handful of additional prompts after clicking “Submit”. Note that using no CV filtering may generate error messages if too many genes are recorded in the data set. Use BRB-filtering to limit the number of genes in your data set, if no CV filtering is used.



Figure 16. The *Filter genes by CV?* field.

The *Minimum sample size per gene in each class* field denotes the minimum number of gene expression values required per gene in each class before calculating gene-gene associations (Figure 17). This field is a text box that only accepts positive whole numbers 3 or higher. Numbers smaller than 3 or any numbers with decimals will generate errors. Please make sure that you do not enter a minimum sample size that is too large. E.g. If your total experiment only has 10 microarray chips, then a minimum sample size of 20 will be impossible. Remember that you cannot compute gene-gene associations if two genes do not share valid expression values on the same chip.



Figure 17. The *Minimum sample size per gene in each class* field.

The *Compute differences between Spearman rank correlations?* field denotes the user's choice to use an approximate statistical test to compare gene-gene Spearman rank correlations between the two classes (Figure 18). Please note that this statistical test may not work well for small sample sizes. If sample sizes are small, you should compare these results against the results from a permutation test. These results will be added to the default output for Fisher's Z-test, which compares gene-gene Pearson correlations between the two classes.

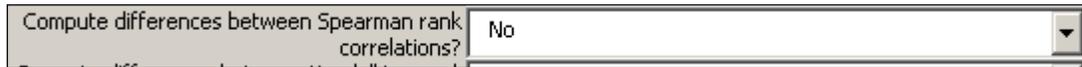


Figure 18. *Compute differences between Spearman rank correlations?* field.

The *Compute differences between Kendall tau rank correlations?* field denotes the user's choice to use an approximate statistical test to compare gene-gene Spearman rank correlations between the two classes (Figure 19). Please note that this statistical test may not work well for small sample sizes. If sample sizes are small, you should compare these results against the results from a permutation test. These results will be added to the default output for Fisher's Z-test, which compares gene-gene Pearson correlations between the two classes.

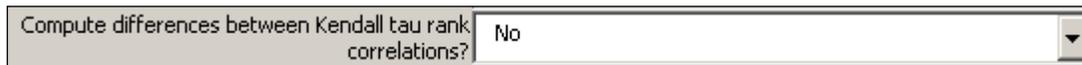


Figure 19. *Compute differences between Kendall tau rank correlations?* field.

The *Choose additional association tests* field allows users to examine and test an optional association measure (Figure 20). By default, the DAPfinder plug-in will always calculate gene-gene Pearson correlations and test for significant differences in gene-gene Pearson correlations with a Fisher's Z test. If users choose *none*, then only the Pearson correlations and Fisher's Z test will be calculated. If users select the option for *mutual information*, *Pearson correlation*, *Spearman rank correlation* or *Kendall rank correlation*, then the additional association will be calculated for all gene-gene pairs and a permutation test will be used to compare the strongest of these gene-gene associations between the two classes. Selecting an additional association test will generate several additional pop-up menu prompts, so you can specify additional testing parameters.

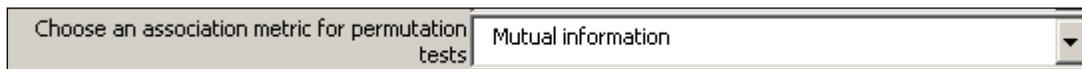


Figure 20. The *Choose additional association tests* field.

The *Display 'top 10' results in HTML output?* field provides user an option to see the "top 10" results from the Fisher's Z test and the "top 10" results from the permutation test (if selected) in the standard HTML report (Figure 21). The default value is FALSE, which removes these "top 10" results from the report. Select TRUE to view the "top 10" results in the HTML report. The HTML report is always generated to display the user

settings for each set of results, but removing the “top 10” results may save a little processing time.



Figure 21. The *Display ‘top 10’ results in HTML output?* field.

The *Enter the name of your saved output data file* field allows users to specify the name of their final results file (Figure 22). The default value is *results.xls*, but note that users can specify any file name with a .xls or .txt file extension. Use a file name with the .txt file extension to avoid losing data in large data sets. Remember MS Excel does not store data sets with more than 65,000 rows. Note, if you run the *DAPfinder* plug-in twice using the same output data file name, the plug-in will generate a pop-up window that asks you if you want to over-write the file or cancel the analysis.

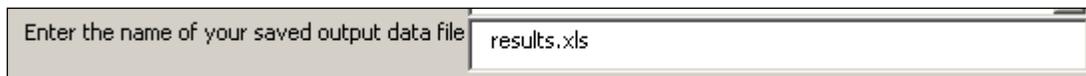


Figure 22. The *Enter the name of your saved output data file* field.

## 7. Specifying Additional Analysis Parameters with Interactive Menus

After you click submit on the *DAPfinder* window, you will most likely see one of several interactive pop-up menus. These pop-up windows are generated by the R script and they are used to prompt the user for additional analysis parameters or alert the user about potential problems. Many of these user prompts used to be included on the *DAPfinder* main window in older versions of the plug-in. Using these interactive pop-up menus yields a cleaner user interface and it ensures users will not enter values for parameters that are not used in the analyses.

Two of the first pop-up windows you may see are the *Overwrite Output File* and *Minimum Sample Size Warning* menus (Figure 23). Before processing your data, the *DAPfinder.R* script searches the file folders in your BRB-ArrayTools project folder to determine if you already have an output file with the same name specified in the *Enter the name of your saved output data file* field of the *DAPfinder* window. If you have already saved an output file with this name, then the *Overwrite Output File* menu (Figure 23, left) will ask if you would like to overwrite the file. Click “OK” to overwrite the file with your new results, or click “Cancel” to cancel the analysis and enter a new name for your output data. The R-script will also check your dataset to make sure it meets the criteria for a successful analysis. If your dataset contains so few microarray chips in one class that it is impossible or unlikely to meet the user specified minimum sample size for any genes, you may see a pop-up message like the *Minimum Sample Size Warning* menu (Figure 23, right). You may click “Cancel” to stop the analysis and choose another data

set, or you may click “OK” to attempt the analysis despite the warning. Other warning messages may indicate problems like too few or too many classes entered, etc.

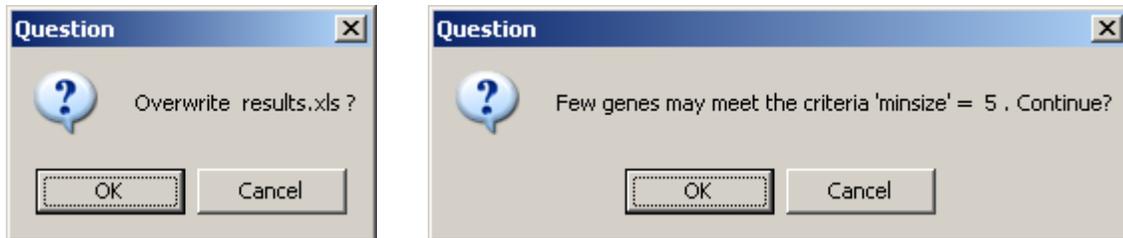


Figure 23. The *Overwrite Output File* and *Minimum Sample Size Warning* menus.

If a user has chosen to upload a list of specific gene-gene pairs for testing, then the plug-in will first prompt the user to select a tab-delimited text file of gene-gene names (Figure 24) and immediately produce an Open file dialog to allow the user to browse for their text file (Figure 25). Note that the Open file dialog will automatically direct you to your current project folder to look for the text file. Remember that the plug-in will apply minimum sample size filtering to all data and it will apply BRB-filtering, outlier removal and CV filtering to the data if specified; therefore, the final output may not contain results for all entered gene-gene pairs. Also, uploading a list of gene-gene pairs will eliminate the gene-gene pair selection choices for users that choose to a second association test and permutation testing. The additional association measure and permutation tests will be computed for all gene-gene pairs in the specified list that pass the filtering criteria. Click “Cancel” on the Open file dialog to stop all analyses.



Figure 24. Prompt message to upload a text file

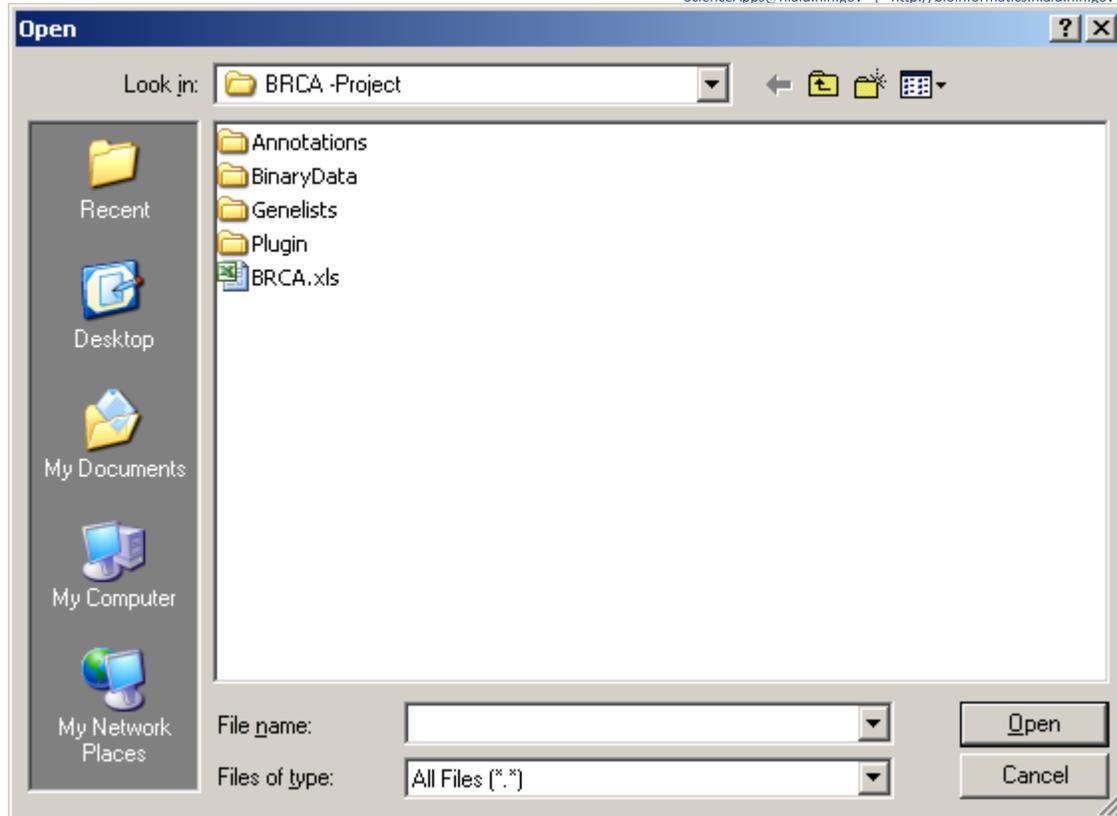


Figure 25. Open file dialog to browse for a text file.

If univariate IQR or SD outlier removal methods were selected, then the next interactive dialog will be the Enter Outlier Removal Constant menu (Figure 26). This menu will allow you to enter an outlier removal constant to customize the outlier removal procedure. The menu will have an editable text box with an appropriate default value for the selected outlier removal method (i.e. IQR default = 1.5, SD default = 4). Choosing a larger constant will remove fewer potential outliers for both methods, while choosing a smaller constant will remove more potential outliers. However, small outlier removal constants may remove too many data points, including non-outliers. Please choose your outlier constant wisely. Decimal values may be entered. Click “OK” to enter a new outlier constant, or click “Cancel” to skip the outlier removal procedure. Note that outlier removal may require a few minutes of computing time for large data sets, so you may need to wait for the next interactive dialog.

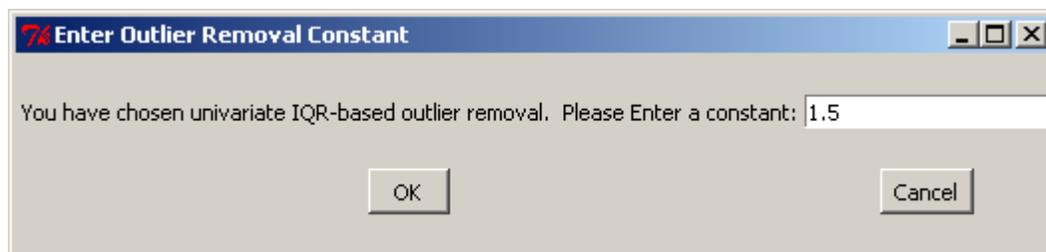


Figure 26. The interactive *Enter Outlier Removal Constant* menu.

If you have chosen to use CV filtering, then next interactive menu is the *Minimum Sample Size Filtering* dialog (Figure 27). It will ask if you would like to apply minimum sample size filtering to your data before applying the CV filter. Remember the plug-in will always apply minimum sample size filtering to ensure that gene-gene associations may be calculated. Minimum sample size filtering will remove genes from the analysis, while the CV filtering process is used to select genes for the analysis. If you apply minimum sample size filtering after CV filtering, then you may end up selecting genes during CV filtering only to remove them moments later. This could dramatically affect sample size. On the other hand, if CV filtering is applied first, then you will not need to do minimum sample size filtering for all genes, only those selected by the CV filter. This may save computational time. The default value is to apply CV filtering before minimum sample size filtering. Choose “Yes” or “No” using the radiobuttons, then click “OK” to apply your choice or click “Cancel” to skip the CV filtering step.



Figure 27. The interactive *Minimum Sample Size Filtering* menu.

After clicking “OK” on the *Minimum Sample Size Filtering* menu above, you will encounter the interactive *Filter by Coefficient of Variation* menu (Figure 28). The menu asks you how many genes should be extracted per class during the CV filtering procedure. The default value is 100. If you select 100 genes, then the CV filtering procedure will select 100 genes with the largest (or smallest) CV from each class, then enter those gene into the association tests. Keep in mind that the CV filtering procedure works independently within each class, so it will not select the exact same genes from each class. Therefore, if you select 100 genes from each class, you may end up with a data set containing 100-200 total genes and you will not be able to predict their exact number. Any whole number may be entered; decimals will likely cause errors. Click “OK” to apply your choice or click “Cancel” to skip the CV filtering step.

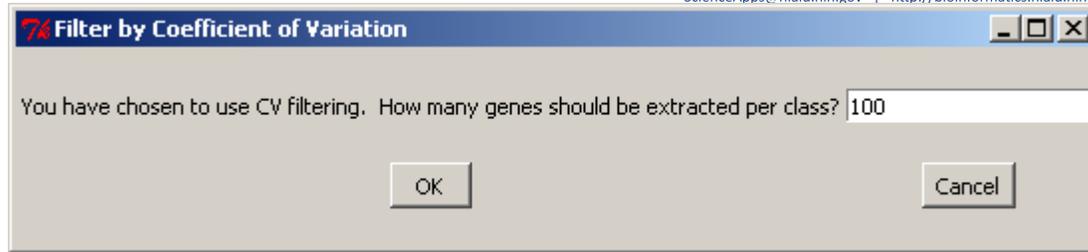


Figure 28. The interactive *Filter by Coefficient of Variation* menu.

After entering all of the settings to select genes for the Fisher's Z test, you will find the interactive *Multiple Testing Adjustments: Fisher's Z test* menu (Figure 29). This radiobutton dialog allows you to choose to display any or all of five available multiple testing adjustments of the Fisher's Z test p-value. If you select "No" for all five choices, then no adjusted p-values will be displayed; if you select "yes" for two or more choices, then only your selected adjustments will be calculated and displayed in the output. Note these multiple testing adjustment choices will also be applied to the approximate test results for Spearman rank correlation and Kendall tau rank correlation, if the user has chosen to report these approximate test results. These choices will not be applied to the permutation test results, but users will have an opportunity to choose a different set of multiple testing adjustments for the permutation test results, if necessary.

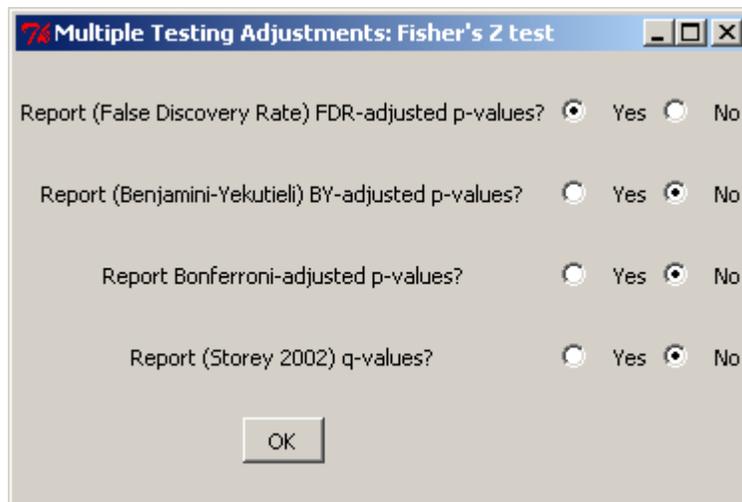


Figure 29. The interactive *Multiple Testing Adjustments: Fisher's Z test* menu.

The first four choices in the *Multiple Testing Adjustments: Fisher's Z test* menu are all adjusted p-value methods (Wright 1992). These adjusted p-values represent the probability of rejecting the null hypothesis after applying a multiple testing adjustment like the Bonferroni correction or False Discovery Rate (FDR) methods. The first two choices of FDR-adjusted p-values and Benjamini-Hochberg (BH) adjusted p-values represent are two variations on the popular and powerful FDR methods frequently used in microarray research (Benjamini and Hochberg 1995). The Benjamini-Yekutieli adjusted

p-values are another variation on the FDR method specifically designed for dependent hypotheses (Benjamini and Yekutieli 2001), which may be very relevant for the multiple correlation tests calculated by this plug-in. The Bonferroni adjusted p-values are based on the well known and highly conservative Bonferroni family-wise error rate adjustment. Adjusted p-values are computed using the `p.adjust()` procedure from the `stats()` package in R.

The final choice in the menu allows you to calculate and display q-values, as described in Storey 2002. These q-values are closely related to the FDR methods cited above. You can interpret each q-value as the minimum FDR that would still produce a significant p-value. So, a  $q = 0.07$  implies that you would have a p-value smaller than 0.05 with the very reasonable  $FDR = 0.07$ , while  $q = 0.78$  implies that in order to produce a p-value smaller than 0.05 you would need to accept an unreasonable  $FDR = 0.78$ . If you were to select all the gene-gene pairs with q-values of 0.10 or lower, it would be safe to assume that only about 10% of the selected genes would be false positives. These q-values are computed using the `qvalues()` package in R (citation). Note the `qvalues()` package requires you to make some assumptions about the number of true positives and false positives in your test p-values. The default methods in the `qvalues()` package will work for many, but not all data sets. When the default `qvalues()` methods do not work, the plug-in will report all other valid results but the column of q-values will not be displayed in your final output.

If you have chosen to calculate an additional association test, you will need to respond to several interactive menus. These interactive dialogs allow you to choose how you will select gene-gene pairs for permutation testing and how you will determine the number of permutations used for each gene-gene pair. However, if you choose to use mutual information as your additional association test, the first prompt you will see is the *Mutual Information Calculation Parameters* menu (Figure 30). This menu allows you to select three different parameters that control how mutual information will be calculated between your gene-gene pairs.

Mutual information is computed using the `build.mim()` and `disc()` procedures of the `minet()` package in R. These procedures allow you to choose one of four different mutual information estimator methods (Gaussian, empirical, Miller-Madow or shrinkage methods), one of two different discretization methods (equal frequencies among bins or equal width bins) and the number of bins used in the discretization function. Previous versions of the plug-in used the default settings of these functions: empirical estimator method, equal frequencies among bins and  $n = \text{floor}(\text{sqrt}(\text{sample size}))$  bins, where sample size denotes the number of microarray chips.

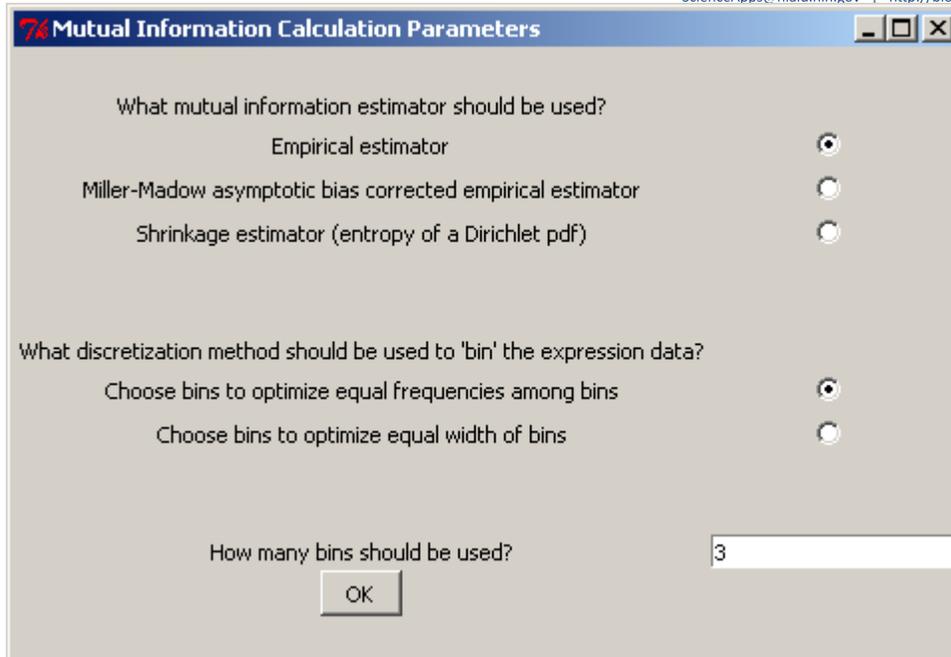


Figure 30. *Mutual Information Calculation Parameters* menu.

Please note that some previous versions of the plug-in had discretized gene expression values separately within each of the two classes. This method could be problematic, because the expression values in the two classes could be discretized into very different types of groups, creating apples to oranges comparisons. The current version of the plug-in discretizes all of the data together, before separating the discretized data into two classes. We believe this method is superior because it creates apples to apples comparisons between the two sets of discretized data.

The next prompt is the *Selection Method for Permutation Tests* menu (Figure 31). It will ask you to choose one of three different gene-gene selection methods to select gene-gene pairs for permutation testing: #1. ‘Top Pairs’ method, #2. Rank-Block-Slice-and-Zoom (RBSZ) method or #3. Select all gene-gene pairs. The default choice is the ‘Top Pairs’ method. Permutation tests require lengthy computation times, so you should only choose the option to *Select all gene-gene pairs* when you have a very small data set. Otherwise, the plug-in may require hours or even days of processing time; it may even crash BRB-ArrayTools. The ‘Top Pairs’ and RBSZ methods both select a subset of genes for permutation testing to reduce computation time. Both methods select genes with the largest differences in gene-gene association between two classes, to ensure that most significant difference in association are identified. Click “OK” to apply your choice or click “Cancel” to cancel the additional association tests.

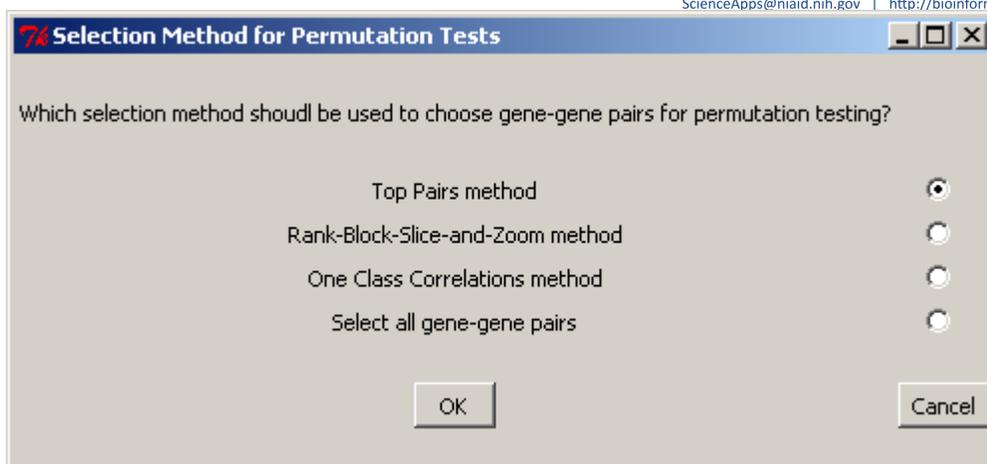


Figure 31. The *Selection Method for Permutation Tests* menu.

The ‘Top Pairs’ selection method identifies  $k$  gene-gene pairs with the largest differences in association between two classes and selects those gene-gene pairs for permutation testing. If the user selects the ‘Top Pairs’ method, then the next interactive dialog will be the *Permutation testing – TOP Gene-gene pairs* menu (Figure 31). This menu asks the user how many of the ‘top’ gene-gene pairs should be selected for the permutation tests. The default value is 20. Users may enter any whole number; decimals may cause errors. Click “OK” to apply your choice or click “Cancel” to cancel the additional association tests.

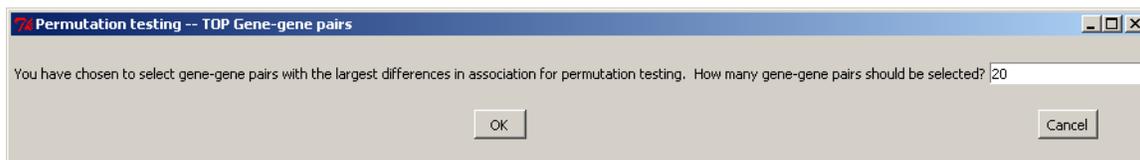


Figure 32. The *Permutation testing – TOP Gene-gene pairs* menu.

The RBSZ method is used to identify most or all of the potentially significant differences in association in a data set without calculating permutation tests for gene-gene pairs that are unlikely to be significant. The RBSZ method has several steps. First, the genes are ranked by the differences in gene-gene association between the two classes. Gene-gene pairs with the largest differences in association are highest ranked. Next, the list of ranked gene-gene pairs is divided into a small number of *blocks*. A small sample, or *slice*, of gene-gene pairs is chosen from each block for some preliminary permutation tests to identify potentially significant gene-gene pairs. The algorithm identifies the lowest ranked block with a significant difference in associations from the preliminary tests, then it “zooms in” on that block to identify the lowest ranked gene-gene pair with a significant difference in association between the two classes. The zoom in takes the selected block and divides it into smaller sub-blocks and sub-slices to find the lowest ranked significant difference in association. After identifying the lowest ranked

significant difference in association, all of the gene-gene pairs ranked higher than this lowest ranked significant gene-gene pair will be selected for complete permutation testing.

If a user chooses the RBSZ selection method, the next interactive dialog will be the *RBSZ Parameters* window (Figure 33). First, the user is prompted to choose the number of blocks used in the RBSZ method. Selecting fewer blocks may speed up computation time, while selecting more blocks may identify more potentially significant gene-gene pairs. Second, users are asked to specify the number of gene-gene pairs per slice in the RBSZ method. Including fewer gene-gene pairs per slice may speed up computation time, while including more gene-gene pairs per slice may identify more potential significant differences in association. Finally, the user may enter the number of preliminary permutations per slice. Fewer permutations will speed up calculations, while calculating more permutations will lead to more reliable identification of the potential significant gene-gene pairs. Click “OK” to apply your choices for the RBSZ parameters or click “Cancel” to cancel the additional association tests.

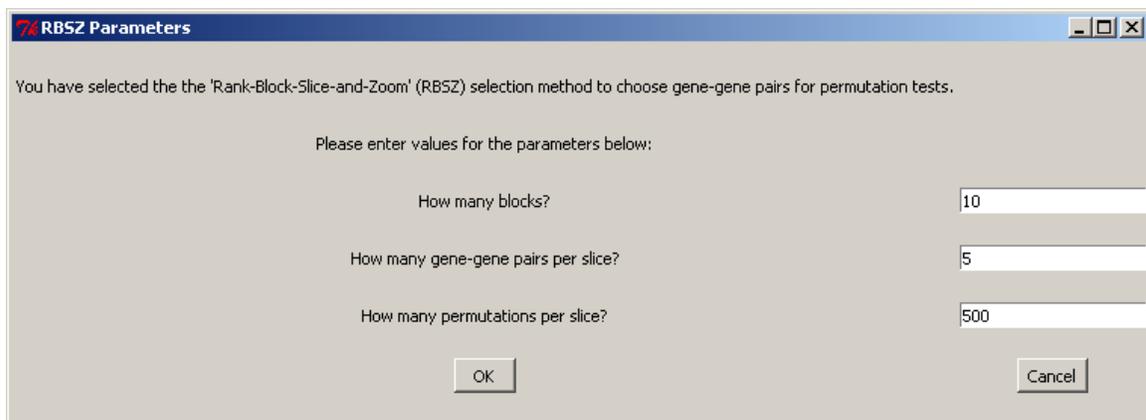


Figure 33. The *RBSZ Parameters* window.

If a user chooses the ‘One Class Correlation’ selection method, then the next interactive dialog will be the *One Class Correlation Parameters* window (Figure 34). First, the user is prompted to choose the p-value threshold used to identify significant gene-gene correlations in each of the two classes. E.g. if the user selects  $p = 0.05$ , then the plug-in will select all gene-gene pairs that have a Pearson correlation with  $p = 0.05$  or less in at least one of the two classes. A smaller p-value threshold will select fewer gene-gene pairs for permutation tests. Second, users are asked to specify the minimum number of gene-gene pairs for permutation testing. Since it is entirely possible that no gene-gene pairs will meet the user’s p-value threshold, the ‘One Class Correlation’ method could choose zero gene-gene pairs for permutation tests. The minimum number of gene-gene pairs field ensures that the user will have some permutation test results. If the user selects a minimum number of gene-gene pairs equal to 20, yet fewer than 20 gene-gene pairs meet the p-value threshold, then the plug-in will select 20 gene-gene pairs with the smallest p-values from each class. Likewise the maximum number of gene-gene pairs

field prevents the ‘One Class Correlation’ method from choosing too many gene-gene pairs for permutation tests. If more gene-gene pairs meet the criteria than the entered maximum value of 100, then the plug-in will choose the 100 gene-gene pairs with the lowest p-values from each class and ignore all other gene-gene pairs meeting the p-value threshold..

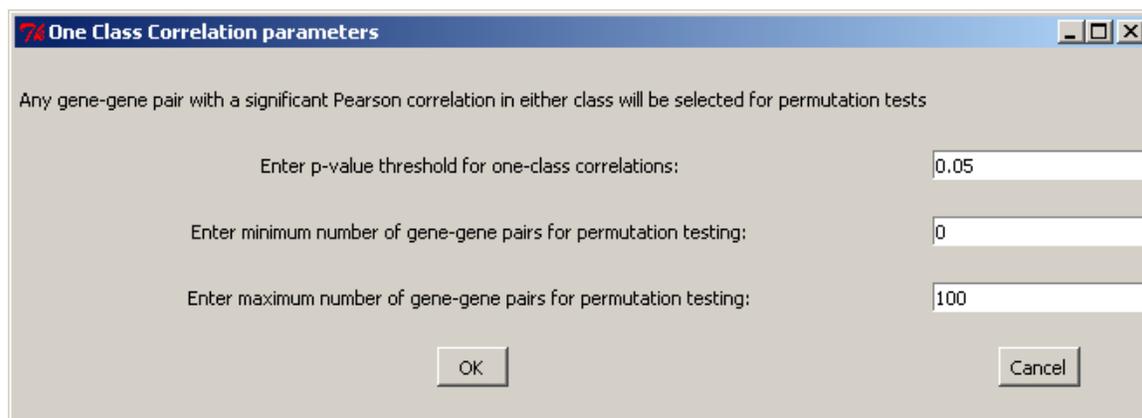


Figure 34. The *One Class Correlation Parameters* window.

After choosing the option to select all gene-gene pairs for permutation testing, or after they enter the parameters for the ‘Top Pairs’, ‘One Class Correlation’ or ‘RBSZ’ selection method, the next interactive window will be the *Permutation Testing Scheme* menu (Figure 35). The user may choose to calculate an equal number of permutations for all selected gene-gene pairs and apply FDR adjustments to the permutation test p-values. Alternatively, users may choose an adaptive permutation scheme that calculates different numbers of permutations for each gene-gene pair. Ideally, the adaptive permutation method will calculate few permutations for differences in association that are obviously highly significant or obviously non-significant, while calculating more permutations for the gene-gene pairs that are nearly significant. This method may reduce computation time, but more importantly it would ensure that precise permutation tests are carried out wherever they are needed. Users must click “OK” to choose a permutation testing scheme. There is no option to cancel at this step.

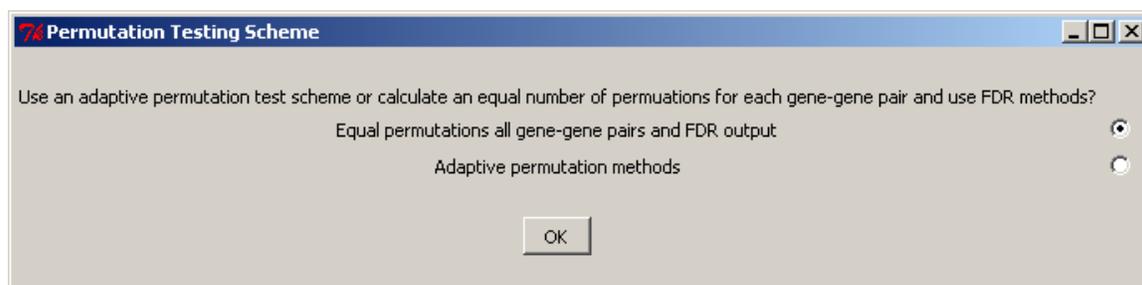


Figure 35. The *Permutation Testing Scheme* menu.

The adaptive permutation method calculates a confidence interval for preliminary permutation p-values to determine the number of additional permutations that should be calculated for each gene-gene pair. This method utilizes the idea that each permutation p-value represents the proportion of permutations more extreme than the true difference in gene-gene association between classes. If each permutation p-value is a proportion, then we can use the binomial distribution to calculate a confidence interval around each permutation p-value (i.e.  $95\% \text{ CI} = p \pm 2 \cdot \sqrt{p(1-p)/n}$ , where  $n$  is the total number of permutations and  $p$  is the permutation p-value). Therefore, if we choose a threshold for the largest significant p-value (e.g.  $p = 0.05$ ) and a confidence level for our binomial confidence intervals (e.g. 95% confidence), we can use these confidence intervals to determine how many permutations should be calculated for each gene-gene pair.

Suppose we have differences in association for three gene-gene pairs and we have calculated 100 permutations for each of these 3 gene-gene pairs. The first gene-gene pair examines a very large difference in association between treated and untreated subjects for genes A and B. Out of 100 permutations, only 1 permutation was more extreme than the true difference in association from these two samples. Therefore the permutation p-value is  $p = 0.01$  and its 95% confidence interval is (0, 0.0299). The 95% confidence interval suggests we can be 95% confident that the true permutation p-value will be between 0 and 0.0299, therefore the gene-gene pair is obviously significant and we do not need to calculate additional permutations for the gene-gene pair A, B. The second gene-gene pair shows a very small difference in association between for genes A and C. We find 84/100 permutations are more extreme than the true difference in association from these two samples, so  $p = 0.84$  and its 95% confidence interval is (0.657616, 1.00). We can be 95% confident the true permutation p-value will be between 0.66 and 1.00, therefore the gene-gene pair is obviously not significant and we do not need additional permutations for gene-gene pair A, C. The third gene-gene pair shows a relatively large difference in association between for genes B and C. We find 6/100 permutations are more extreme than the true difference in association from these two samples, so  $p = 0.06$  and its 95% confidence interval is (0.011256, 0.108744). This confidence interval overlaps our threshold p-value of 0.05, therefore we need to compute more permutations to determine if this difference in associations is statistically significant.

If the user selects the adaptive permutation method, the next pop-up menu will be the *Adaptive Permutation Test Parameters* window (Figure 36). The menu will allow users to enter a threshold value for the largest significant p-value and confidence level for the binomial confidence intervals used in the adaptive permutation method. A smaller threshold value for the largest significant p-value (e.g.  $p = 0.0001$ ) will identify fewer significant gene-gene pairs and may reduce computation times, while larger threshold values for the largest significant p-value (e.g.  $p = 0.05$  or  $p = 0.10$ ) will identify more significant gene-gene pairs but may require more computation time. The default threshold for the largest significant p-value is  $p = 0.01$ . Lower confidence levels (e.g. 0.95) will decrease computation time, but may lead to less precise identifications. Higher confidence levels (e.g. 0.999) will require more computation time, but may be more precise. The default confidence level is 0.999. Click “OK” to apply your choices for the

adaptive permutation test parameters or click “Cancel” to cancel the additional association tests.

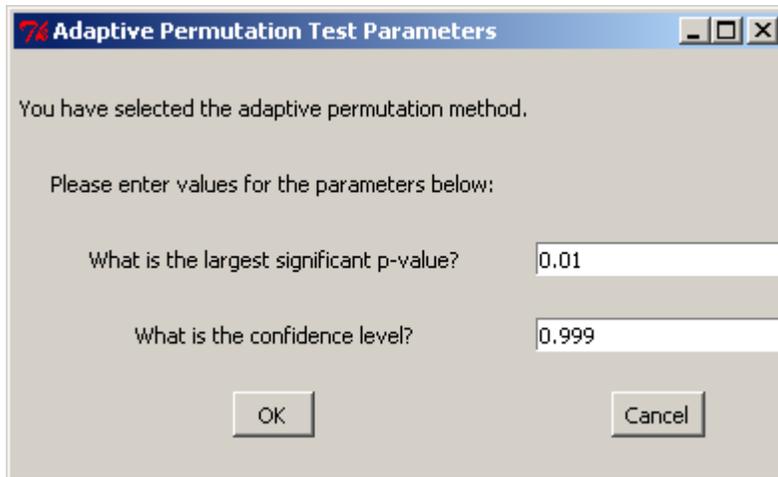


Figure 36. The *Adaptive Permutation Test Parameters* window.

If the user selects equal permutation tests for all gene-gene pairs, then the next interactive pop-up menu will be the *Equal Permutations* window (Figure 37). The menu will allow the user to specify the number of permutations calculated for each gene-gene pair. Fewer permutations will result in faster computations, while more permutations will produce more precise permutation p-values. For example, p-values calculated for 100 permutations would only be accurate to two decimal places (e.g.  $p = 4/100 = 0.04$ ). It is important to note that p-values calculated from a very small number of permutations will be very unreliable (e.g. if only 10 permutations are used, p-values will be very unstable), but extremely large numbers of permutations will not add much useful precision to the p-values (e.g. 100,000 permutations vs. 10,000 permutations will provide accuracy to 6 decimal places vs. 5 decimal places, but it will not affect the reliability of the permutation p-values). Click “OK” to apply your choice for the number of permutations or click “Cancel” to cancel the additional association tests.

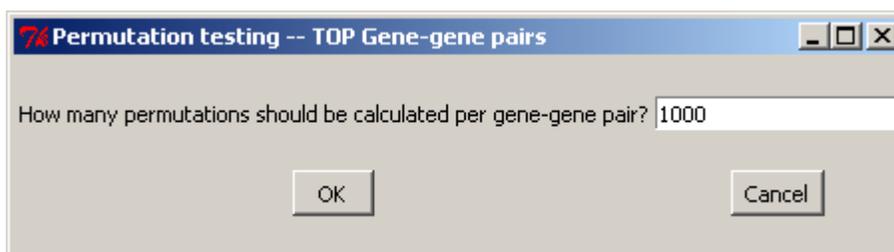


Figure 37. The *Equal Permutations* window.

If the equal permutations method is selected, you will first see a warning message to remind you of the possible consequences of applying multiple testing corrections to data after first selecting a subset of all hypothesis tests from the entire family of tests (Figure 38). When you select a subset of hypotheses from a list of all possible tests, you could introduce some dependencies among the selected hypotheses. Also, since you will likely only want to select the most significant hypotheses for further testing, you will probably have difficulties trying to estimate the true proportion of null hypotheses among your tests (i.e. the number of true negatives). For these reasons, you should think carefully about whether or not you want to apply multiple testing adjustments to the permutation tests.

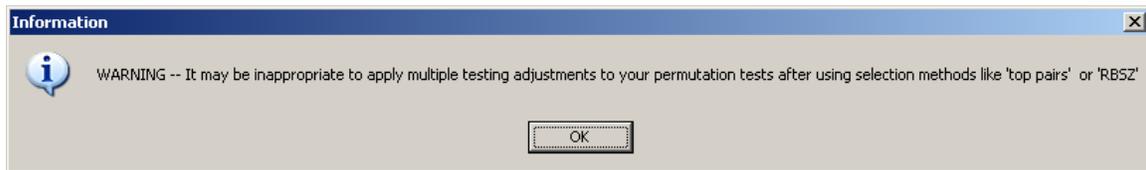


Figure 38. Warning about multiple testing adjustments.

After the warning, you will see the *Multiple Testing Adjustments: permutation test* menu (Figure 39). This dialog is identical to the *Multiple Testing Adjustments: Fisher's Z test* menu, except that it now applies correction methods to the permutation test results. If you would like to heed the previous warning, remember that you can answer "No" to all five choices to avoid any adjustments. Also, like with the Fisher's Z test, it is possible that the default settings for the `qvalues()` procedure will not work for the permutation test. If this is the case, then no q-values will be displayed. The `qvalues()` procedure in R may be more likely to fail for the permutation test, because there will be few large p-values calculated and it will be more difficult to estimate the true proportion of null hypotheses for the permutation tests than for the Fisher's Z-tests.

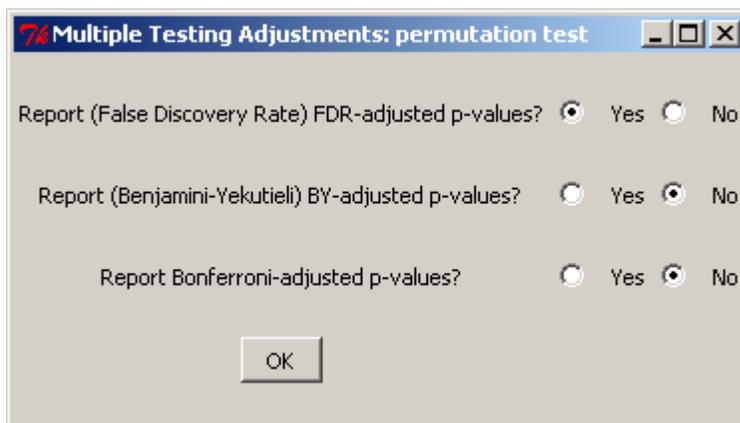


Figure 39. The *Multiple Testing Adjustments: permutation test* menu

After specifying the input values in the last window, the plug-in will process your data and complete all of your tests. When the plug-in is finished, your web browser will automatically open an HTML report which records all of your settings and several diagnostic statements. Remember that you will not see all of the interactive pop-up windows described above. You will only see the pop-up windows relevant to the options you have selected in the *DAPfinder* window or previous interactive pop-up menus. Look for new pop-up windows and watch for the final HTML report to appear. Computations will be delayed until all the interactive pop-up menus have been answered.

## 8. Interpreting the output

The *DAPfinder* plug-in produces two pieces of output, an HTML report and an output data file in MS Excel or tab-delimited text file format. The HTML report contains one or two tables describing the user inputs to the plug-in (Figure 40). This allows you to verify and record the user settings for each set of output. The results may contain additional tables of the “top 10” results for Fisher’s Z test and permutation tests, if the option is selected. Finally, the HTML report will contain some log messages about empty columns or rows in the data set, the number of gene-gene associations calculated, the number of outliers removed, the number of permutation tests performed, the required processing time in R and the location of the stored output data file.

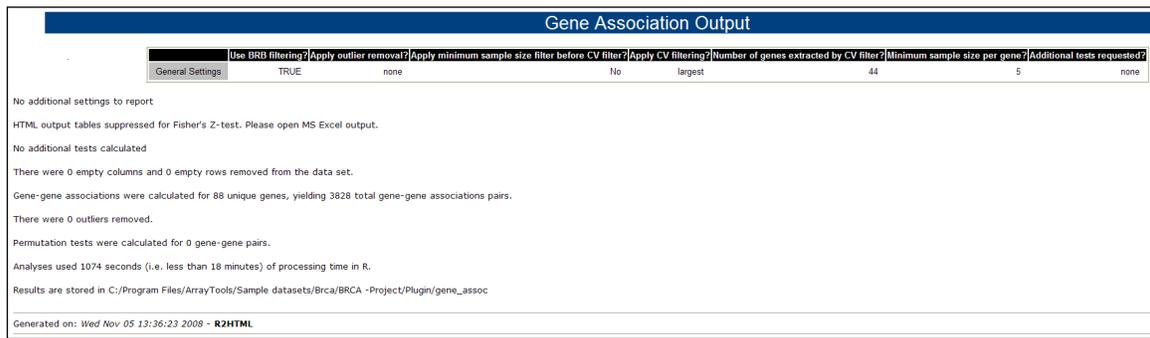


Figure 40. HTML Report with one table of user inputs.

The output data file can be opened in MS Excel. It will contain all the necessary statistical and descriptive results, including the gene names, Pearson correlation values, Fisher’s Z test statistic, p-values, etc (Figure 41). These results will be sorted by p-values and other measures to ensure that only the most relevant results are displayed in Excel.

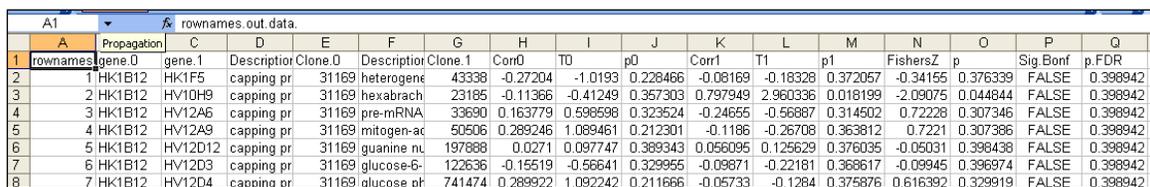


Figure 41. The output data file.

## 8. Literature Cited

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289–300.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165–1188.

Storey JD. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**: 479-498.

Wright, S. P. (1992). Adjusted P-values for simultaneous inference. *Biometrics*, **48**, 1005–1013.

## 9. Summary

Please contact [ScienceApps@niaid.nih.gov](mailto:ScienceApps@niaid.nih.gov) with any questions concerning the *DAPfinder* plug-in. This manual describes the *DAPfinder* plug-in beta version 0.1 and its features. User input is always appreciated.

Jeff Skinner, M.S  
Biostatistics Specialist  
Bioinformatics and Computational Biosciences Branch (BCBB)  
NIH / NIAID / OD / OSMO / OCICB  
[ScienceApps@niaid.nih.gov](mailto:ScienceApps@niaid.nih.gov)  
<http://bioinformatics.niaid.nih.gov>