Introduction of MAAMD Workflow

Section A. Environment Configuration

This section introduces how to set up a computer to run MAAMD. Proper software need be set up before the run of MAAMD. This configuration is a one-time installation procedure.

Required software

- Kepler 2.4
- AltAnalyze 2.0.8
- JDK 7 or above
- R 3.0.0 or above

Software Installation

1. Installation of R
   - Download R from http://cran.r-project.org/ and install
   - Add the folder path where R.exe locates to the 'system variables' list.
     For Windows,
     - Right click "Computer" and select "Properties"
     - Go to "Advanced system settings", click the sub-menu "Advanced"
     - Select "Environment Variables…"
     - Scroll down the "System Variables" list and select the variable "Path"
     - Add the R.exe path to the end of the path, separate it from the existing variables using ";".
       - For 64-bit OS, the path should be like "C:\Program Files\R\R-3.0.0\bin\x64"
       - For 32-bit OS, the path should be like "C:\Program Files\R\R-3.0.0\bin\i386".
2. Installation of Kepler
   - Go to http://www.oracle.com/technetwork/java/javase/downloads/index.html to download and install JDK.
   - Go to https://kepler-project.org/users/downloads to download and install Kepler
   - To check whether Kepler is installed properly, start Kepler by double clicking its icon. You should see Kepler's graphical user interface. If a dialog for updating modules appears, select 'yes'; Kepler will automatically restart after installing those updated modules.
   - After installation, go to the directory where Kepler is installed, open file /common-2.4.0/resources/configurations/configuration.xml. Modify the line "<maxWaitTime>300</maxWaitTime>" to "<maxWaitTime>-1</maxWaitTime>", this will allow the pop-up webpages to wait until the user makes a decision.

Note: Once you update your Kepler, the corresponding configuration.xml need be updated due to the setting of Kepler. The updated configuration.xml can find at $HOME/KeplerData/kepler.modules/common-2.4.X/resources/configurations/.

3. Installation of AltAnalyze

- Go to http://code.google.com/p/altanalyze/downloads/list?can=1&q and download v.2.0.8.
- Unzip to your desired directory.
- To make sure AltAnalyze works properly and has installed the species database, go to the command-line console; for windows, go to "start", select "run" and type "cmd".
  - Change the directory to the location where AltAnalyze.py is installed using a command line like "cd C:\tools\AltAnalyze_v.2.0.8".
  - Type "AltAnalyze.exe", you should see AltAnalyze's graphical user interface if AltAnalyze can start properly. Otherwise, please check the version of your AltAnalyze.
  - Click "Begin Analysis"; a prompt window will appear which indicates no species database found if this is the first time you've run AltAnalyze. Click "Continue" and select the species which you want to analyze, then click "Continue". AltAnalyze will download corresponding resources automatically. Click "Quit" after the downloading is complete.

4. Installation of R packages

- Make sure that you have the permission to update R libraries.
- For windows, go to the directory where R is installed, such as "C:\Program Files\R\R-3.0.0", right-click the folder and select "properties". Under "security" tab, edit the permission and make sure you have the 'write' permission.
- Double-click R shortcut to open R console. If both 32-bit and 64-bit are installed, pick the correct one for your OS.
- Input the following commands in R console to install bioconductor packages:

```
source("http://www.bioconductor.org/biocLite.R")
biocLite()
biocLite("affyQCReport")
biocLite("GEOquery")
biocLite("arrayQualityMetrics")
```

- Input the following commands to test whether the packages have been installed successfully

```
library(affyQCReport)
library(GEOquery)
library(arrayQualityMetrics)
```

You need install these libraries properly before you run MAAMD workflow.

Section B. MAAMD Procedure

This section introduces how to use MAAMD and prepare input files.

1. Download MAAMD zip package, and unzip to C:, so you will have a folder "C:\MAAMD" which contains a "workflow" folder and a "sample" folder

2. Search GEO database http://www.ncbi.nlm.nih.gov/geo/. Look for data sets and collect data set information.
   Edit input CSV files for the selected data sets with the fixed file format. Refer to "datasets.csv" for the format of the summary of datasets and "datainfo-gse9400.csv" and "datainfo-gse33100.csv" for the format of the samples in an individual data set.

Note: please do not modify the names of columns.
Note: The suffix ".CEL" is required for both "SampleName" and "NewName".

3. Run Kepler and open MAAMD workflow in Kepler, keep Internet connection open when MAAMD is running.

4. Edit the parameters for MAAMD. Note: all paths have to use forward slash, namely "/", for path delimiter. "\" does not work in Kepler.
   Nset: the number of datasets that you want to analyze.
   Note: If Nset is smaller than what you listed in datasets.csv, then MAAMD will analzye the first "N" data sets only.
   WorkPath: the folder where you want to store the data and results.
   DataFile: the path of the csv file where you collect all datasets' information.
   Note: This csv file contains the summary of all targeted datasets. Don't assign the path to those csv files for sample information in individual data sets. The path of sample information file is assigned in this csv. Refer to "datasets.csv" as an example.
   MAAMDPath: the folder where you store MAAMD workflows.
   Note: homologene.txt must be stored in the same folder as a reference file.
   AltAnalyze: the directory of AltAnalyze location, for example, "C:/AltAnalyze_v.2.0.8-Win64".

5. Click "run" button

Note: the workflow has tested in Windows and Mac systems.

MAAMD Workflow Selection

1. **Combined workflow: MAAMD-Download-Analysis-Comparison.xml**
   This workflow combines all steps together, including data downloading, data analysis and result comparison. The combined workflow runs all sets as a batch in order, so the users simply follow the workflow. The combined workflow generates folders to store data and results in the workflow and hence can locate required data automatically. This will avoid potential mistakes such as the mismatch of targeted data location.

2. **Discrete workflow : MAAMD-Download.xml, MAAMD-Analysis.xml and MAAMD-Comparison.xml**
   In the main, these three workflows perform the same functions as the combined workflow. The discrete workflows allow you executing them one by one and hence are more flexible to handle.
   MAAMD-Download.xml downloads and decompresses targeted data sets.
   MAAMD-Analysis.xml analyzes the downloaded data sets
   MAAMD-Comparison.xml compares the results of analyzed data sets.

   The advantages of using the discrete workflow are:

a. The user can control the process more flexibly. For example, the user can compare the results in different ways by running MAAMD-Comparison.xml several times.
b. It gives the users more options for data analysis. For example, some of the authors named their sample files following their own rules, which were different from the recommended methods (ie: 'GSM23510.CEL'). If the sample names are not correctly assigned, the subsequent workflow can't find the data files and errors will result. With the discrete workflow, you can download all your data sets, check them, and run MAAMD-Analysis.xml using datainfo.csv with your unique sample names.
c. The discrete workflow allows the analysis of the existing data. Since the discrete workflow separates the data acquisition step from the analysis step, it is not necessary for the data to be available online. MAAMD-Analysis.xml can analyze any Affymetrix data, stored locally or acquired online. Subsequently, MAAMS-Comparison.xml can compare any AltAnalyze results stored in your computer.

With the discrete workflow, you must ensure that the data or results are at the assigned location before you want to use them, and that the data folders have the correct structure.

Section C. Workflow Processes

The following describes what you'll observe when you run the combined workflow MAAMD-Download-Analysis-Comparison.xml:

1. MAAMD creates folders in your WorkPath named by the assigned "Dataset" in your datasets.csv.

2. MAAMD downloads and uncompresses all targeted datasets, the downloading progress is indicated by a progress bar.
   Folders containing CEL files can be observed after this step.

3. MAAMD analyzes individual data sets one by one.
   a. CEL files in the targeted folder are renamed with your assigned names.
   b. A webpage pops up and asks you to select CEL files for quality control.
   Note: some data sets contain CEL files acquired by different platforms. You need to select files acquired by the same platform.
   c. A "temp" folder is created. All selected files are copied to this "temp" folder.
   d. A "report" folder is created in the 'temp' folder to store quality-control results.
   e. A webpage pops up to show you the results of quality control analysis. In the "report" folder, many plots have been created. You can copy the "report" folder to save it elsewhere, as the "temp" folder will be deleted later.
   f. A webpage pops up and asks you to select CEL files for AltAnalyze analysis based on the results of quality control analysis.
   Note: All selected CEL files must be acquired by the same platform.
   g. A "result" folder is created in the folder for the current data set. Sub-folders such as ExpressionInput, ATP-output are created.

h. All selected CEL files are copied into the "result" folder.

i. A "groups.txt" file is created in the "result" folder including all selected data.

j. A webpage pops up and asks you to select the comparisons between groups.

k. A "comps.txt" file is created in the "result" folder containing all selected comparisons.

l. AltAnalyze runs and several more folders, such as "ExpressionOutput" and "GO-Elite" are created.

m. AltAnalyze exports the results in to the "result" folder.

n. A Kepler display pops up with the summary of the execution of AltAnalyze.

4. A "comparison.html" file is created in the assigned WorkPath.

5. A webpage pops up and asks which data sets you want to compare.

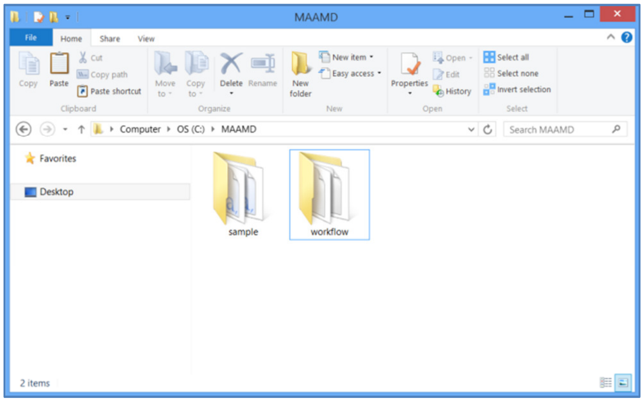
6. A Kepler display opens and reminders you that MAAMD is comparing your selected dataset.

7. A Kepler display pops up to tell you whether homologous genes have been found across data sets. If homologous genes are found across selected datasets, a file named "ComparisonSets.txt" can be found in the WorkPath folder.

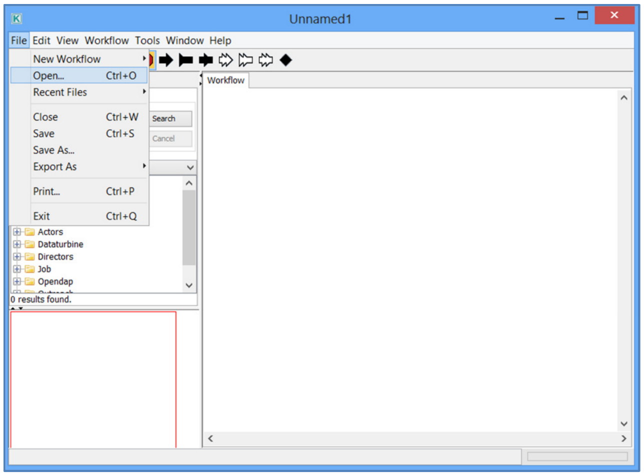
Section D. A Test Case (Windows)

This section shows an example how to use MAAMD to analyze datasets of interest. The required input files are available as supplementary files.

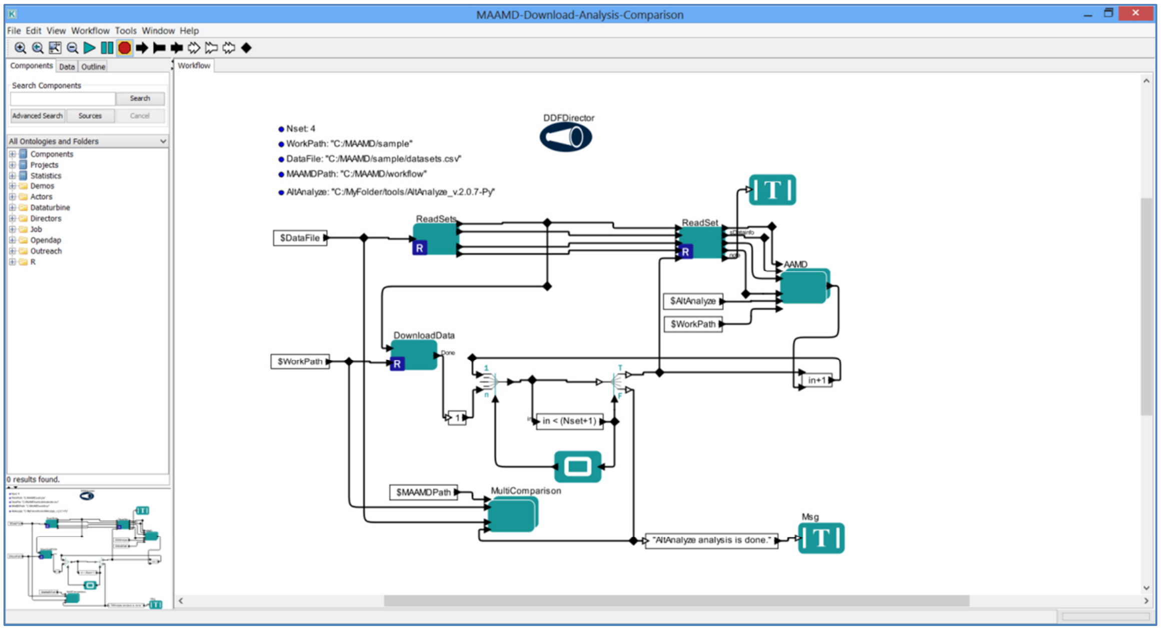
1. Make sure that you have the unzipped MAAMD package at "C:\MAAMD"



2. Double click "Kepler" icon to start Kepler as the figure below shows.

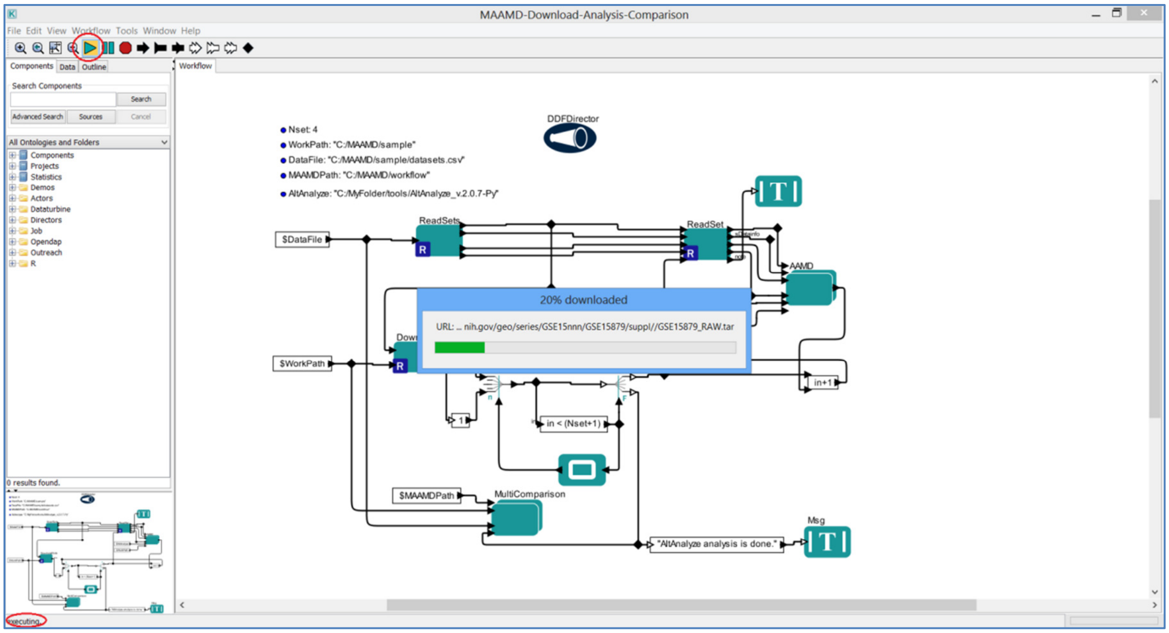


3. Open ‘MAAMD-ALL.xml’ under "C:\MAAMD\workflow".



- Nset: 4
- WorkPath: "C:/MAAMD/sample"
- DataFile: "C:/MAAMD/sample/datasets.csv"
- MAAMDPath: "C:/MAAMD/workflow"
- AltAnalyze: "C:/MyFolder/tools/AltAnalyze_v.2.0.7-Py"

4. Change the value of "AltAnalyze" as the directory of AltAnalyze in your OS and then click "run" button at the top. MAAMD starts with "executing" prompt at the bottom of Kepler. A processing bar will pop up to show the downloading progress. Sometimes, the bar hides behind other windows. The required time depends on the Internet speed and the size of data sets.



MAAMD-Download-Analysis-Comparison

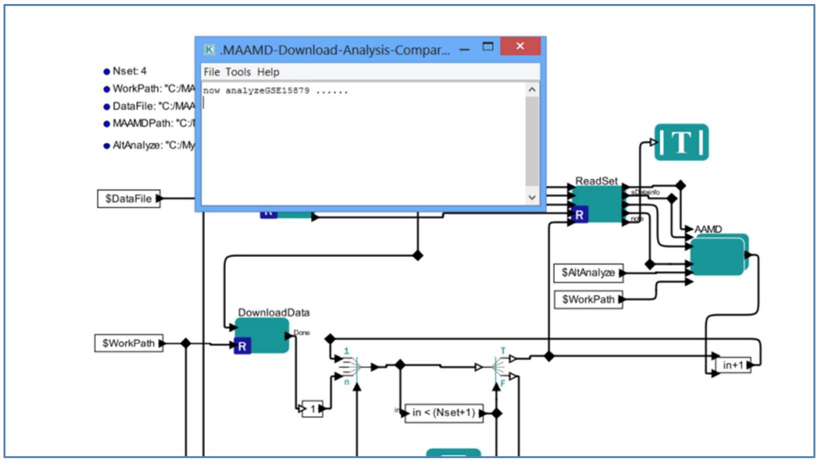
- Nset: 4
- WorkPath: "C:/MAAMD/sample"
- DataFile: "C:/MAAMD/sample/datasets.csv"
- MAAMDPath: "C:/MAAMD/workflow"
- AltAnalyze: "C:/MyFolder/tools/AltAnalyze_v.2.0.7-Py"

20% downloaded

URL: ... nih.gov/geo/series/GSE15nnn/GSE15879/suppl//GSE15879_RAW.tar

"AltAnalyze analysis is done."

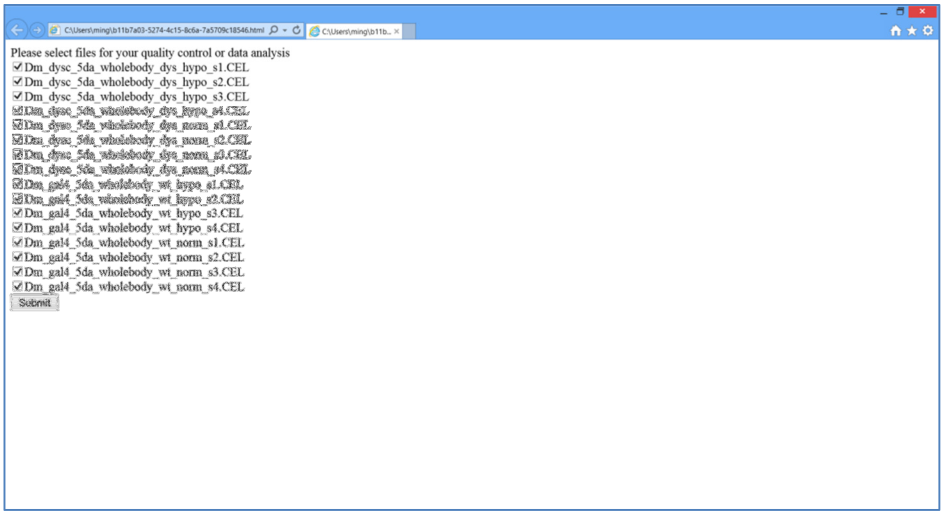
5. After the download is completed, MAAMD starts to analyze downloaded data. A prompt window pops up as a reminder of the analysis progress.



.MAAMD-Download-Analysis-Compar...

```
now analyzeGSE15879 ......
```

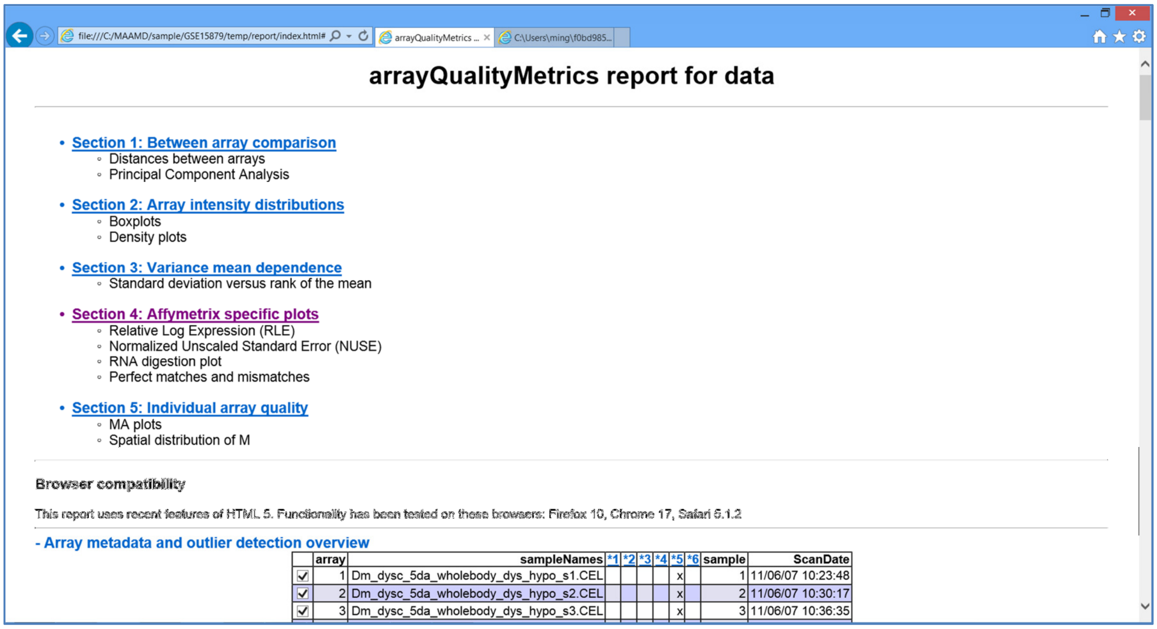
6. In the meantime, a webpage opens with a list of the samples in the first data set. This webpage allows you to pick the CEL files for quality control. It is recommend that you select all samples if these samples were acquired using the same platform. If they were acquired by different platforms, please select only the samples acquired by the same platform for one analysis. The required time depends on the number of selected files.



Please select files for your quality control or data analysis

☑ Dm_dysc_5da_wholebody_dys_hypo_s1.CEL
☑ Dm_dysc_5da_wholebody_dys_hypo_s2.CEL
☑ Dm_dysc_5da_wholebody_dys_hypo_s3.CEL
☑ Dm_dysc_5da_wholebody_dys_hypo_s4.CEL
☑ Dm_dysc_5da_wholebody_dys_norm_s1.CEL
☑ Dm_dysc_5da_wholebody_dys_norm_s2.CEL
☑ Dm_dysc_5da_wholebody_dys_norm_s3.CEL
☑ Dm_dysc_5da_wholebody_dys_norm_s4.CEL
☑ Dm_gal4_5da_wholebody_wt_hypo_s1.CEL
☑ Dm_gal4_5da_wholebody_wt_hypo_s2.CEL
☑ Dm_gal4_5da_wholebody_wt_hypo_s3.CEL
☑ Dm_gal4_5da_wholebody_wt_hypo_s4.CEL
☑ Dm_gal4_5da_wholebody_wt_norm_s1.CEL
☑ Dm_gal4_5da_wholebody_wt_norm_s2.CEL
☑ Dm_gal4_5da_wholebody_wt_norm_s3.CEL
☑ Dm_gal4_5da_wholebody_wt_norm_s4.CEL

7. After quality control is complete, a quality report window appears, as below. You can read the report and decide which samples will be qualified for further analysis.



arrayQualityMetrics report for data

- Section 1: Between array comparison
  - Distances between arrays
  - Principal Component Analysis
- Section 2: Array intensity distributions
  - Boxplots
  - Density plots
- Section 3: Variance mean dependence
  - Standard deviation versus rank of the mean
- Section 4: Affymetrix specific plots
  - Relative Log Expression (RLE)
  - Normalized Unscaled Standard Error (NUSE)
  - RNA digestion plot
  - Perfect matches and mismatches
- Section 5: Individual array quality
  - MA plots
  - Spatial distribution of M

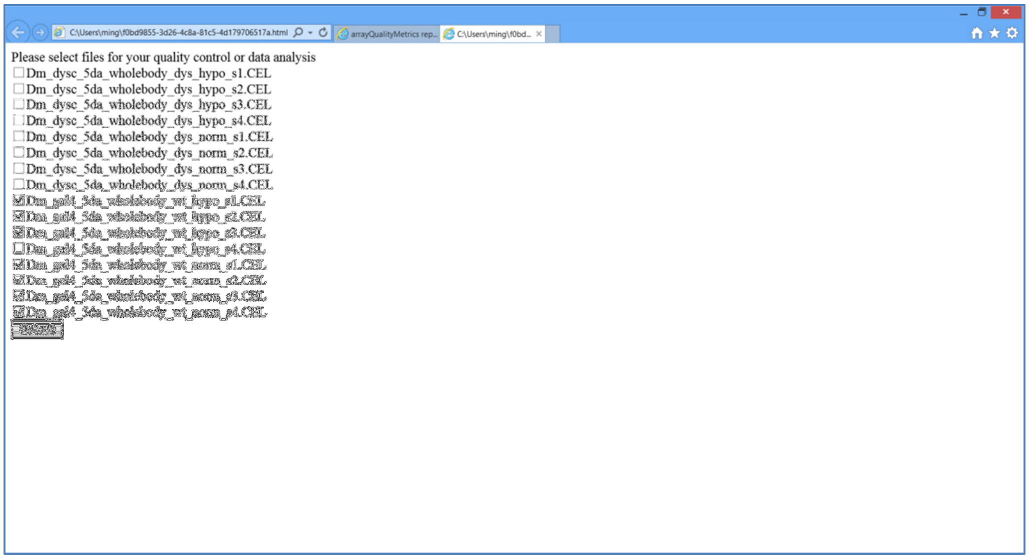
Browser compatibility

This report uses recent features of HTML 5. Functionality has been tested on these browsers: Firefox 10, Chrome 17, Safari 5.1.2

- Array metadata and outlier detection overview

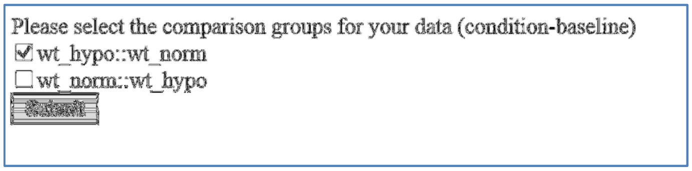
	array	sampleNames	*1	*2	*3	*4	*5	*6	sample	ScanDate
☑	1	Dm_dysc_5da_wholebody_dys_hypo_s1.CEL					x		1	11/06/07 10:23:48
☑	2	Dm_dysc_5da_wholebody_dys_hypo_s2.CEL					x		2	11/06/07 10:30:17
☑	3	Dm_dysc_5da_wholebody_dys_hypo_s3.CEL					x		3	11/06/07 10:36:35

8. A webpage with the list of samples pops up again. This time, this webpage allows you to select samples for microarray analysis. So you can exclude samples with poor quality. In the figure below, those samples from "dysc" flies were not selected in order to compare only wild type flies.


C:\Users\ming\f0bd9855-3d26-4c8a-81c5-4d179706517a.html
arrayQualityMetrics rep...
C:\Users\ming\f0bd...

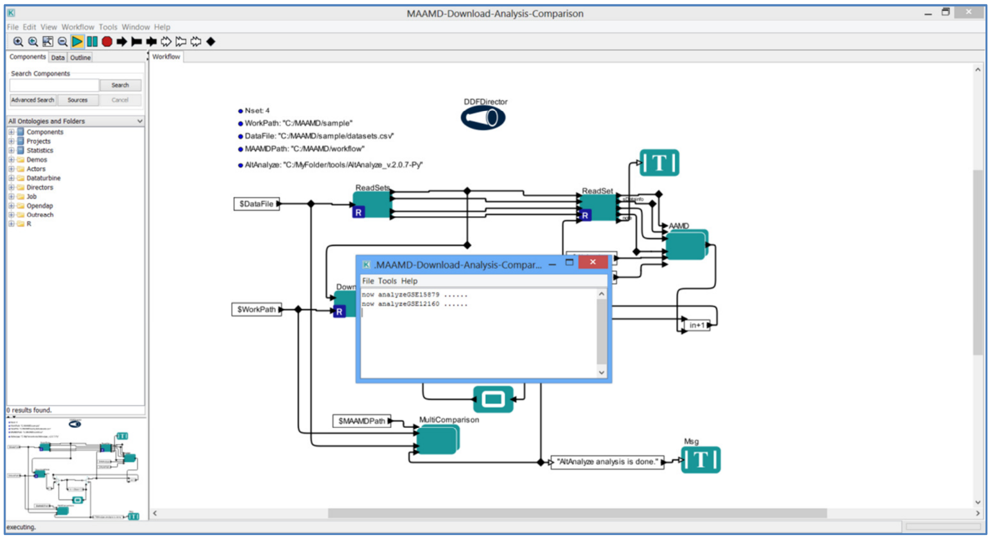
Please select files for your quality control or data analysis
☐ Dm_dysc_5da_wholebody_dys_hypo_s1.CEL
☐ Dm_dysc_5da_wholebody_dys_hypo_s2.CEL
☐ Dm_dysc_5da_wholebody_dys_hypo_s3.CEL
☐ Dm_dysc_5da_wholebody_dys_hypo_s4.CEL
☐ Dm_dysc_5da_wholebody_dys_norm_s1.CEL
☐ Dm_dysc_5da_wholebody_dys_norm_s2.CEL
☐ Dm_dysc_5da_wholebody_dys_norm_s3.CEL
☐ Dm_dysc_5da_wholebody_dys_norm_s4.CEL
☑ Dm_gal4_5da_wholebody_wt_hypo_s1.CEL
☑ Dm_gal4_5da_wholebody_wt_hypo_s2.CEL
☑ Dm_gal4_5da_wholebody_wt_hypo_s3.CEL
☐ Dm_gal4_5da_wholebody_wt_hypo_s4.CEL
☑ Dm_gal4_5da_wholebody_wt_norm_s1.CEL
☑ Dm_gal4_5da_wholebody_wt_norm_s2.CEL
☑ Dm_gal4_5da_wholebody_wt_norm_s3.CEL
☑ Dm_gal4_5da_wholebody_wt_norm_s4.CEL

9. Based on your selected samples, MAAMD will find the groups for these samples and make a webpage that allows you to select your desired comparisons. The comparisons are paired; the latter one is the baseline condition. For example, the first row in the figure below shows wt_norm as the baseline condition and the second row shows wt_hypo as the baseline. Please select only one between these two matched rows.
   At least one comparison shall be selected for further analyses.



Please select the comparison groups for your data (condition-baseline)
☑ wt_hypo::wt_norm
☐ wt_norm::wt_hypo

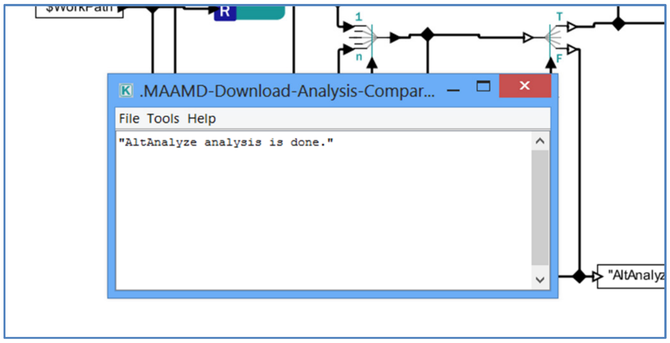
10. MAAMD calls AltAnalyze to analyze your selected samples. AltAnalyze will run in the background, so you just need to keep Kepler running. The required time depends on the number and species of your samples. Samples with large genome sizes take longer
11. After MAAMD finishes the analysis of the first data set, the analysis-progress prompt window updates as showed below. The same steps shown in steps 6 - 10 are repeated for the left data sets.



- Nset: 4
- WorkPath: "C:/MAAMD/sample"
- DataFile: "C:/MAAMD/sample/datasets.csv"
- MAAMDPath: "C:/MAAMD/workflow"
- AltAnalyze: "C:/MyFolder/tools/AltAnalyze_v.2.0.7-Py"

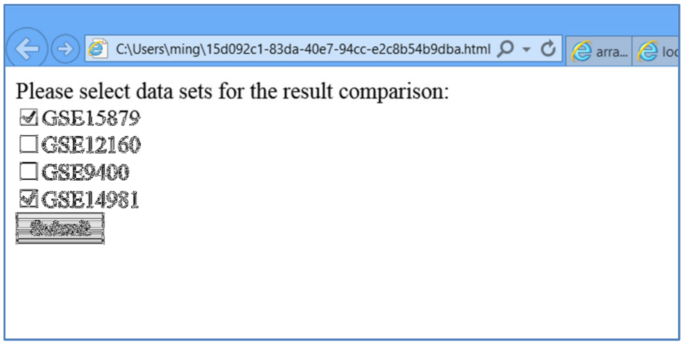
```
now analyzeGSE15879 ......
now analyzeGSE12160 ......
```

12. After AltAnalyze analyzes all targeted data sets, a prompt window pops up announcing the completion of AltAnalyze.



```
"AltAnalyze analysis is done."
```

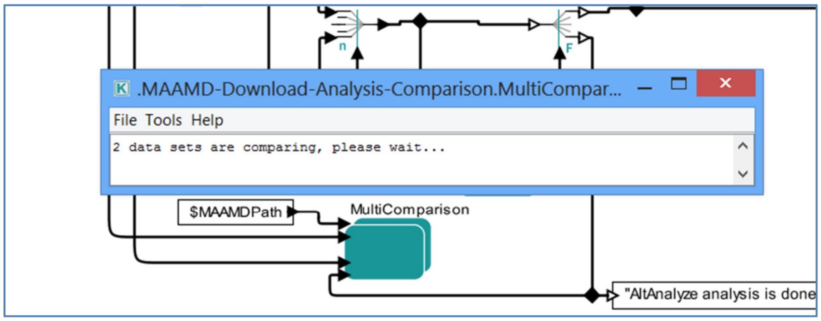
13. Then MAAMD starts to make a comparison between analyzed results to acquire the conserved homologous genes in different data sets. A webpage pops up to allow the user to select the targeted data sets. You can select all data sets, but remember: the more you selected, the fewer conserved genes are expected. If you want to compare them in different ways, you may run MAAMD-comparison.xml independently after you finish the whole process.



Please select data sets for the result comparison:

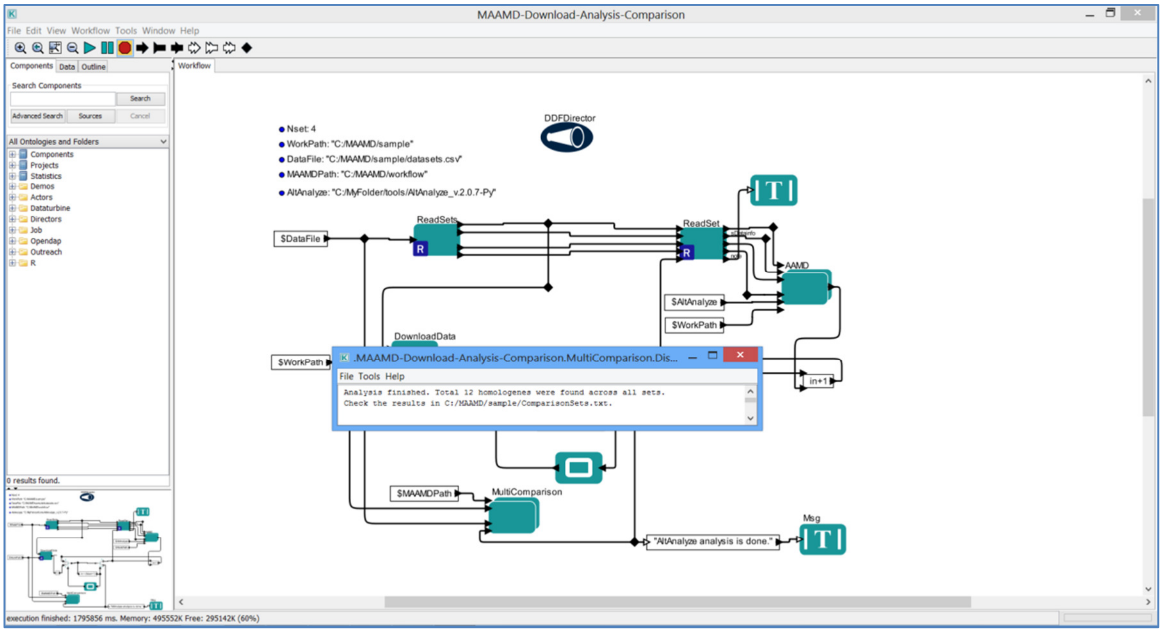
☑ GSE15879
□ GSE12160
□ GSE9400
☑ GSE14981

14. This prompt window reminds you that you are comparing two data sets.


n
F
.MAAMD-Download-Analysis-Comparison.MultiCompar...
File Tools Help
$MAAMDPath
MultiComparison
"AltAnalyze analysis is done

```
2 data sets are comparing, please wait...
```

15. After the comparison is complete, a prompt window pops up to inform you of the location where the comparison results are stored.



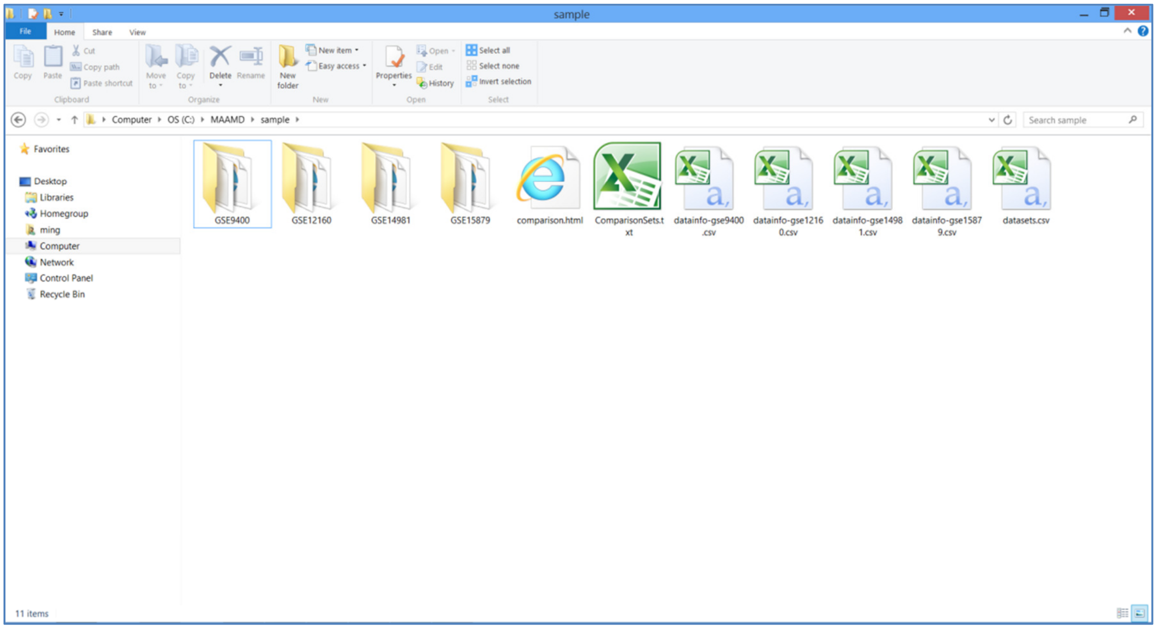
- Nset: 4
- WorkPath: "C:/MAAMD/sample"
- DataFile: "C:/MAAMD/sample/datasets.csv"
- MAAMDPath: "C:/MAAMD/workflow"
- AltAnalyze: "C:/MyFolder/tools/AltAnalyze_v.2.0.7-Py"

.MAAMD-Download-Analysis-Comparison.MultiComparison.Dis...

```
Analysis finished. Total 12 homologenes were found across all sets.
Check the results in C:/MAAMD/sample/ComparisonSets.txt.
```

```
execution finished: 1795856 ms. Memory: 495552K Free: 295142K (60%)
```

16. The figure below shows the final outputs of MAAMD under "C:\MAAMD\sample".



sample
File Home Share View
Copy Paste Cut Copy path Paste shortcut
Clipboard
Move to Copy to Delete Rename
Organize
New folder New item Easy access
New
Properties Open Edit History
Open
Select all Select none Invert selection
Select
Computer OS (C:) MAAMD sample
Search sample
Favorites
Desktop
Libraries
Homegroup
ming
Computer
Network
Control Panel
Recycle Bin
11 items

GSE9400
GSE12160
GSE14981
GSE15879
comparison.html
ComparisonSets.txt
datainfo-gse9400.csv
datainfo-gse12160.csv
datainfo-gse14981.csv
datainfo-gse15879.csv
datasets.csv