How to setup and use the MAYDAY Affymetrix Processing pipeline (AffyPP)

Anna Jasper

Preparations

Essential prerequisites

The MAYDAY Affymetrix Processing Pipeline (AffyPP) is part of the MAYDAY Processing Framework (MPF). As it includes an R-function to normalize the data, the R Interpreter with its dependencies and the BioConductor package has to be installed first. The statistical modules used by the pipeline require the `apache.commons.math` library. The `.jar`-file is available at the MAYDAY homepage `www.zbit.uni-tuebingen.de/pas/mayday/download/download.html`.

Installing R and the BioConductor package

R can be downloaded from `http://cran.r-project.org/`.
After installing R, the BioConductor package needs to be installed. In order to do so open your R command window and type the following:

```
source("http://bioconductor.org/biocLite.R")
biocLite()
```

1 Using the Affymetrix processing Pipeline (AffyPP)

1.1 Start the pipeline and select cel-files

To start the pipeline go to:

$$DataSet \rightarrow Importfromfile$$

A filechooser opens (see figure 1).

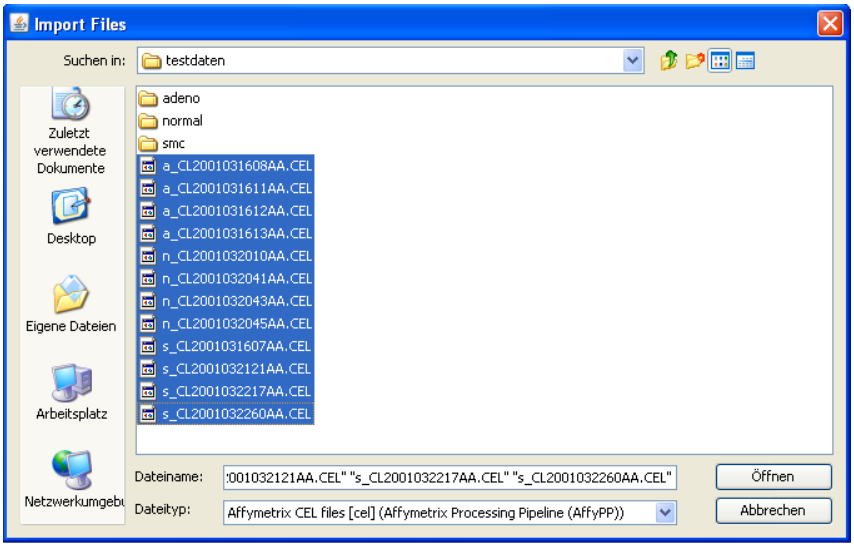


Figure 1: filechooser for selecting the cel-files

Select

"Affymetrix Cel files [cel] (Affymetrix Processing pipeline (AffyPP))"

as file type and then select all your cel-files and press *"OK"*.

1.2 Specify options for AffyPP

The MPF-Applicator opens (see figure 2) and you have to specify the following options:

1. **Files:** shows the number of files you have selected. Your selection can be changed by pressing "'browse"'
2. **Groups:** you don't have to specify any groups now. A class selection dialog will open after the normalization of the data for assigning groups
3. specify a name for the MIO groups which later contain the p-values. The individual MIO group names will be set to "user entered MIO name + ANOVA" for the ANOVA p-value and "user chosen MIO name + post-hoc t-test group n vs. m". The MIO group names are also important to for the p-value correction methods.

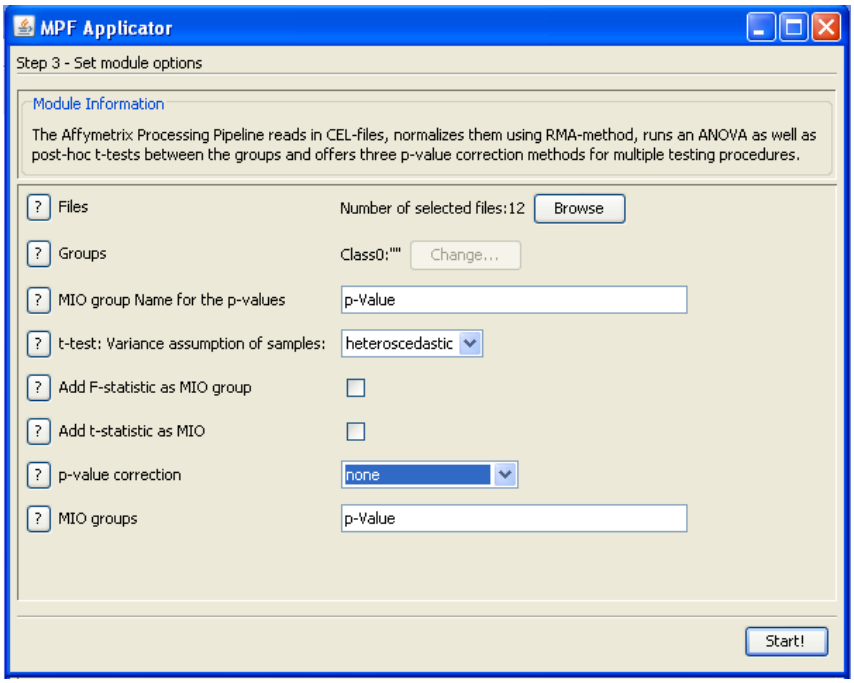


Figure 2: Specify options for AffyPP

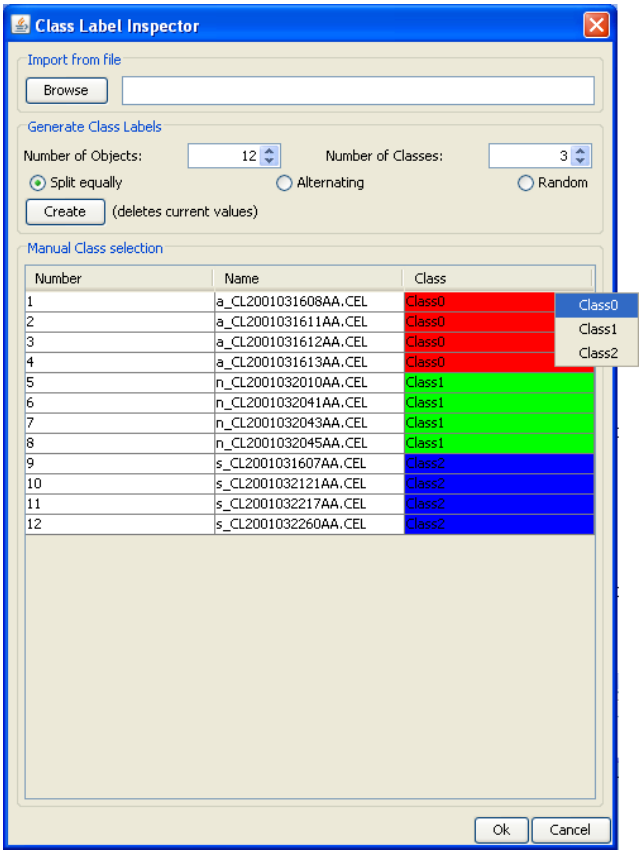
4. **t-test variance assumption of samples:** You have to choose whether you assume a heteroscedastic or homscedastic variance
5. select this if you want to have the F-statistic corresponding to the ANOVA p-value
6. select this if you want to have the t-statistic for every post-hoc t-test
7. select which p-value correction method you want to apply:
   - none
   - Bonferroni
   - Holm's Stepwise-Correction
   - False Discovery Rate (FDR)
8. enter the names of the MIO groups containing the p-values. If you can't remember them, leave it empty. The pipeline then opens a dialog, where all MIO groups are shown and you can select them easily.

In the example a heteroscedastic variance was assumed and no p-value correction was chosen. So the p-value correction method was set to *"none"*.

1.3 Assign classes

After the data is normalized a class selection dialog opens, where you can assign each cel-file to a class. To do so, choose the number of classes. The classes can

either be automaticly assigned (equally splitted or alternating) or manually by a right click on the class column and selection of the class. The example consists



of three classes. So the number of classes was set to *"3"*, and automaticly split (*"split equally"* → *"create"*) was chosen. The files were assigned as following (see figure 1.3):

class 1	**class 2**	**class 3**
a_CL2001031608AA.CEL	n_CL2001032010AA.CEL	s_CL2001031607AA.CEL
a_CL2001031611AA.CEL	n_CL2001032041AA.CEL	s_CL2001032121AA.CEL
a_CL2001031612AA.CEL	n_CL2001032043AA.CEL	s_CL2001032217AA.CEL
a_CL2001031613AA.CEL	n_CL2001032045AA.CEL	s_CL2001032260AA.CEL

1.4 Filter p-values

When the analysis is finished a MIO-Group-Selection dialog opens. Here you can select the p-values which should be used for filtering. The filter criteria is set to < 0.05. If multiple groups are selected, the criteria will be ALL p-values < 0.05. In

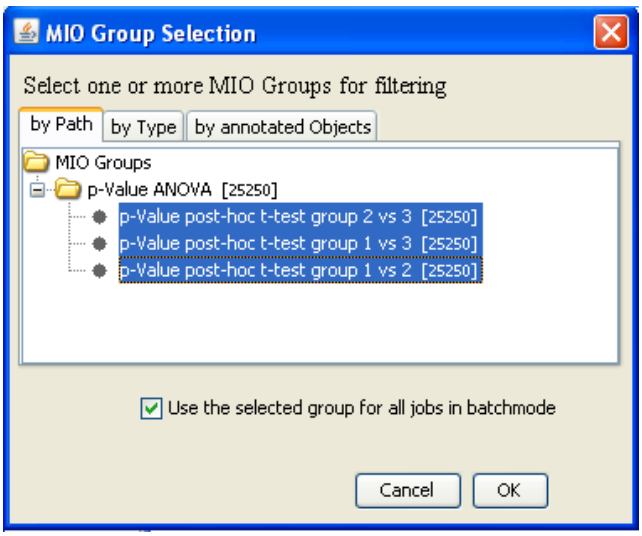


Figure 3: MIO Group Selection Dialog

the example all post-hoc t-test p-values (see figure 3) were chosen.

1.5 Set dataset properties and return data to MAYDAY

After the filtering is done a windows opens, were the dataset properties can be viewed and changed (see figure 4).

When pressing *"OK"* the dataset is returned to MAYDAY. A global probe list containing all probe sets (12625 probe sets) as well a filtered probe list (57 probe sets) containing all probe sets, which matched the criteria (here a p-value < 0.05 in ALL post-hoc t-tests) is returned to MAYDAY (see figure 5).

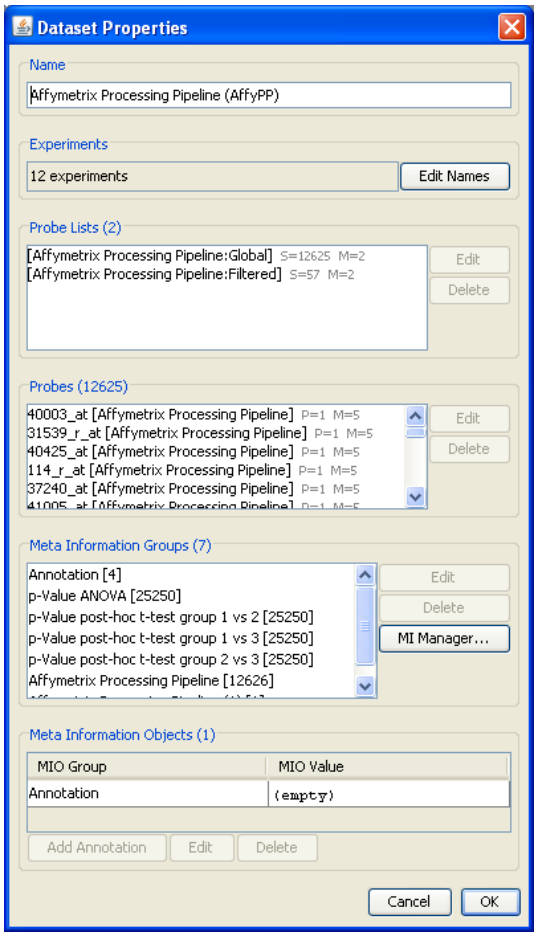


Figure 4: set dataset properties

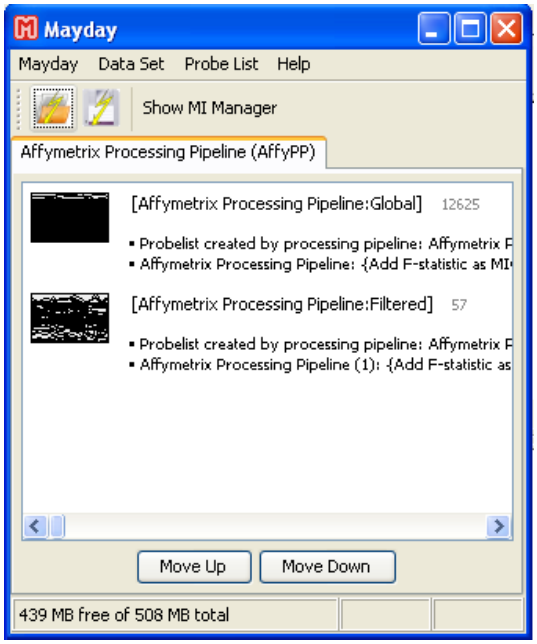


Figure 5: data returned to MAYDAY