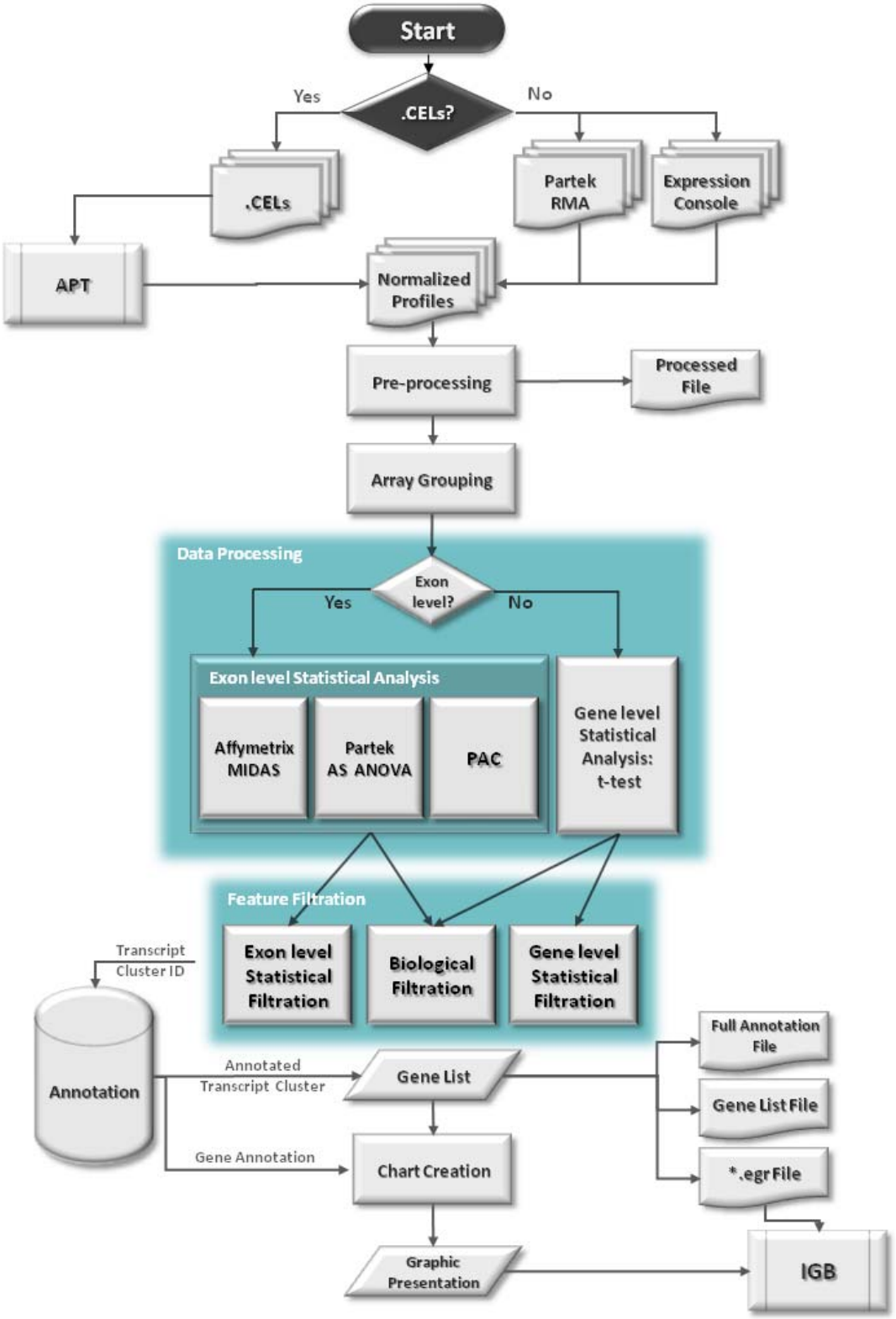
Tutorial for the easyExon

Version 1.0.3

Contents

0. Flowchart of easyExon



1. Alt-Splicing Study

Step 0: Input Data Format Selection

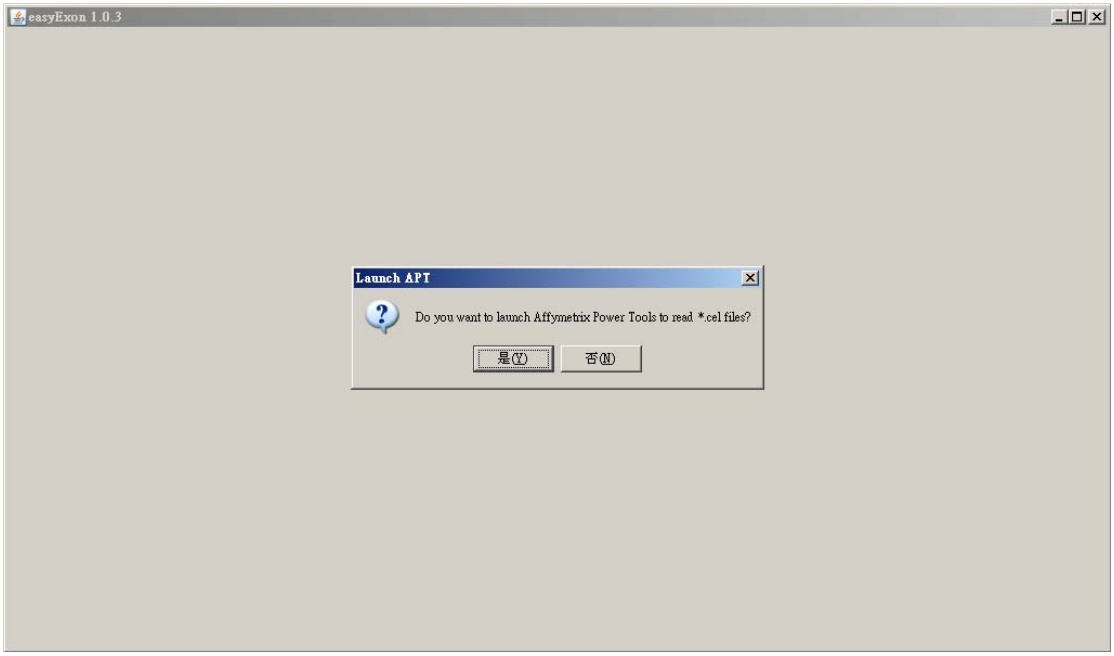
User needs to select the input data format. Two data formats are allowed:

(i) CEL files

Choose **YES** to launch Affymetrix Power Tools (APT), and follow the **Step 1.a**.

(ii) *.txt summary files

Choose **NO** to load *.txt summary files, and follow the **Step 1.b**.



Step 1: Data Preparation

Step 1.a: Get summary files by Affymetrix Power Tools (APT)

Before using easyExon to analyze exon arrays, user have to prepare Affymetrix Power Tools, corresponding library files, and *.CEL files. When these files are ready and corresponding directory paths are set, easyExon will launch APT to get the summary file that contains signals of each exon across all samples.

Data preparation:

(i) Affymetrix Power Tools (APT)
APT can be downloaded from the website:
http://www.affymetrix.com/support/developer/powertools/index.affx

(ii) Library files
Library files for APT have to be in the same directory. Those library files can be downloaded from the website:
http://www.affymetrix.com/support/technical/libraryfilesmain.affx

Here are the descriptions of library files.

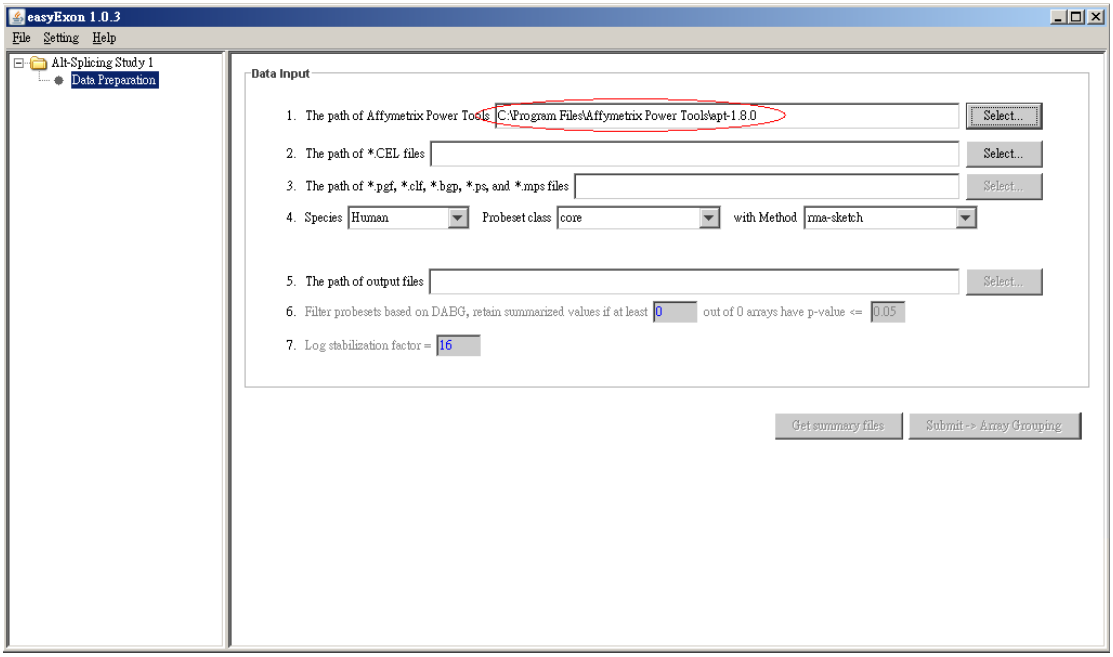
Extension	Content
*.pgf	File defining probe sets **E.g.:** HuEx-1_0-st-v2.r2.pgf
*.clf	File defining x, y <-> probe id conversion **E.g.:** HuEx-1_0-st-v2.r2.clf
*.bgp	File defining probes to be used for GC background **E.g.:** HuEx-1_0-st-v2.r2.antigenomic.bgp
*.ps	File specifying probe sets to summarize **E.g.:** HuEx-1_0-st-v2.r2.dt1.hg18.core.ps
*.mps	File containing meta probeset definitions. File must contain a probeset_id column and a probeset_list column. **E.g.:** HuEx-1_0-st-v2.r2.dt1.hg18.extended.mps

(iii) Probe results files (CEL files)
Put all CEL files in the same directory. The CEL files in the same directory will be used to get summary files.

Directory path setting:

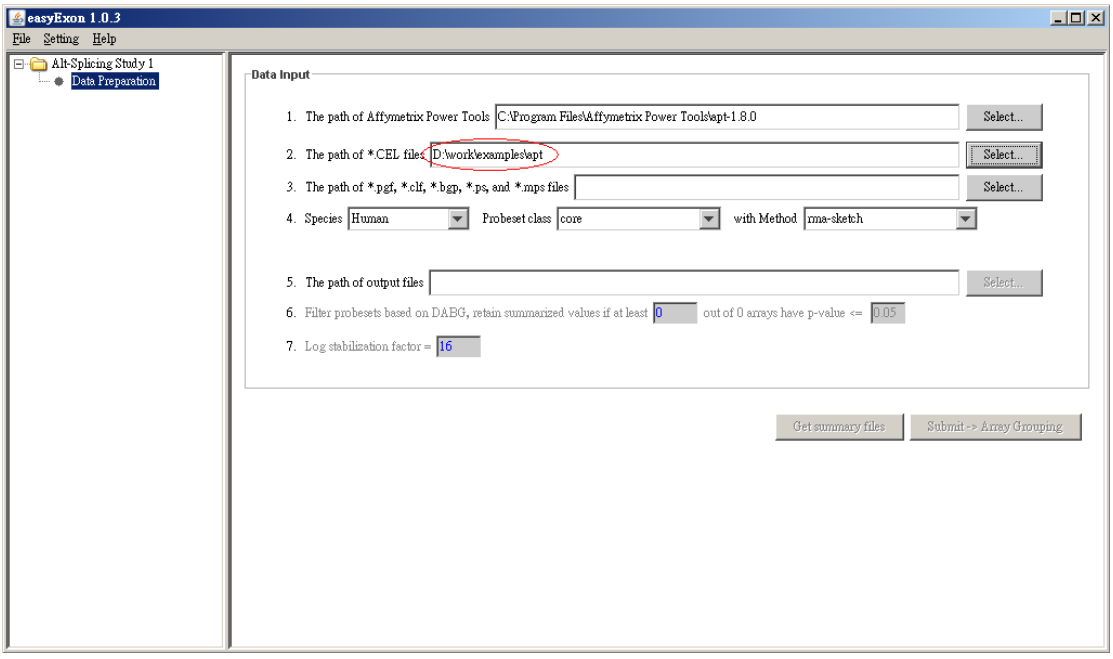
(i) Path of the APT

Select the path of APT which contains the “bin” directory.



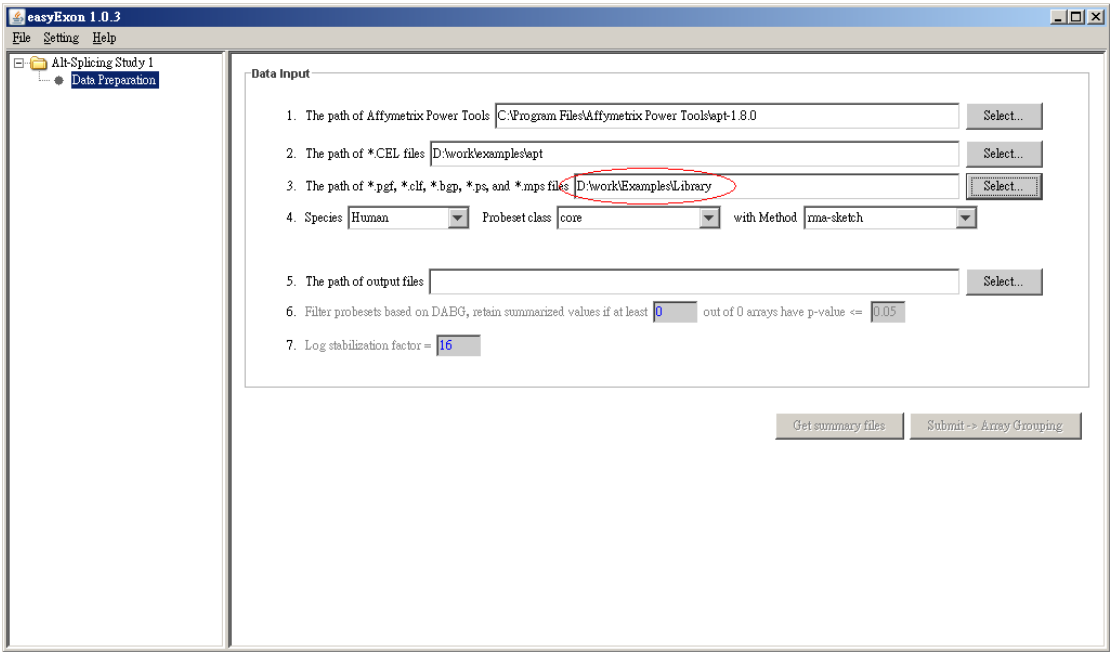
(ii) Path of CEL files

Select the directory which includes CEL files.



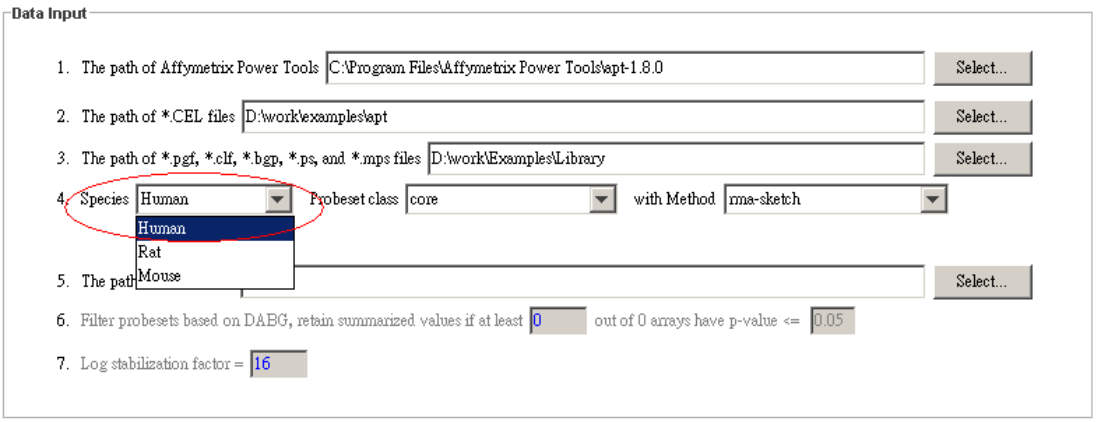
(iii) Path of library files

Select the path of library files directory which includes *.pgf files, *.clf files, *.bgp files, *.ps files, and *.mps files.

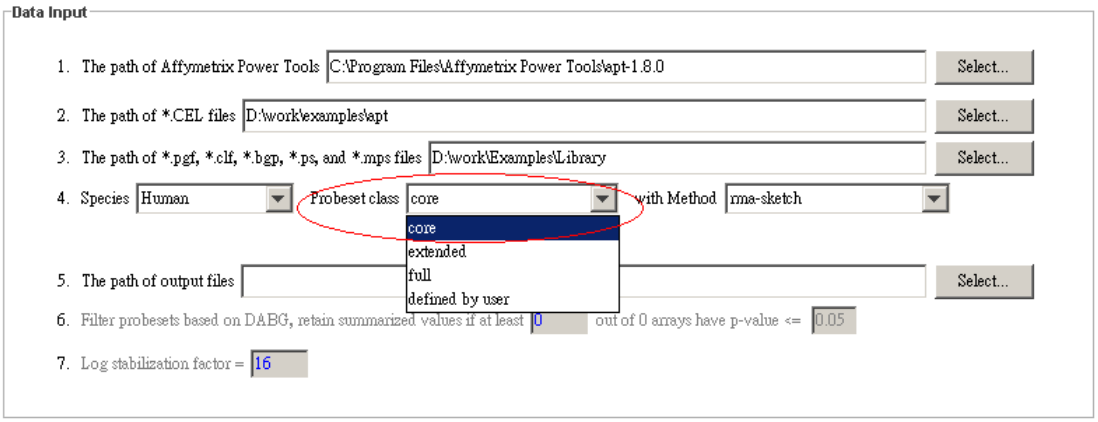


(iv) Select the species, probeset class and the method for getting summary files.

Species selection (Human, Rat, or Mouse)



Probeset class selection (core, full, extended, or defined by user)



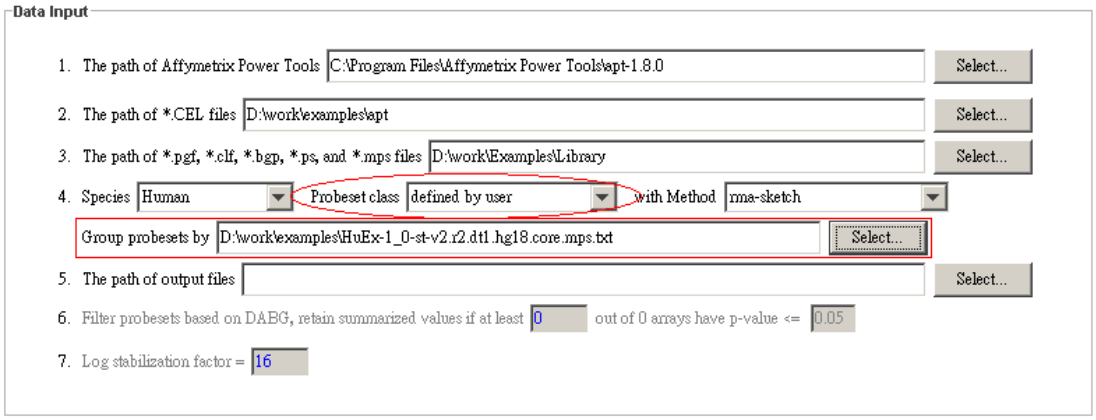
When user selects “defined by user,” user needs to select the meta-probesets file location.

Meta-probesets file format:

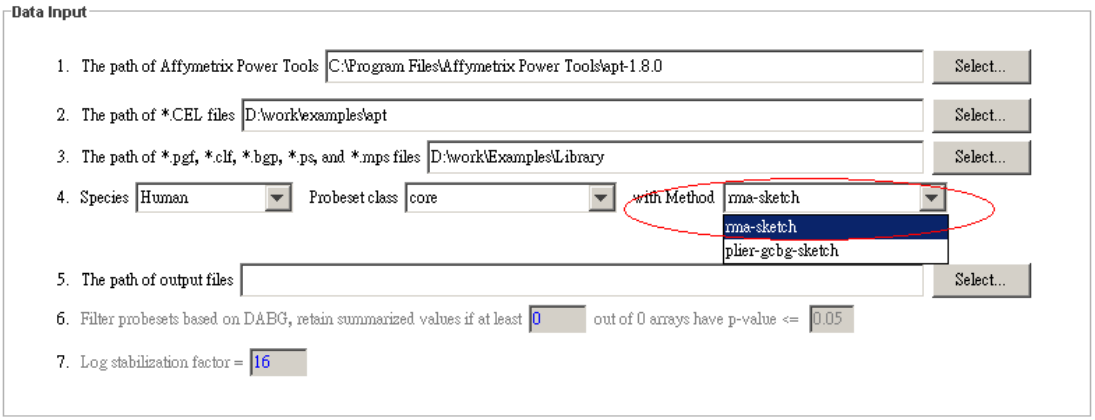
The First row of the data contains “probeset_id”, “transcript_cluster_id”, and “probeset_list”. The first column contains probeset_ids; the second column is transcript_cluster_id. The third column is probeset_list with corresponding transcript_cluster_id.

E.g.:

```
probeset_id<Tab>transcript_cluster_id<Tab>probeset_list<Tab>…
3096575<Tab>3096575<Tab>3096630 3096631 3096632<Tab>…
2949118<Tab>2949118<Tab>2949119 2949120 2949121 2949130<Tab>…
```

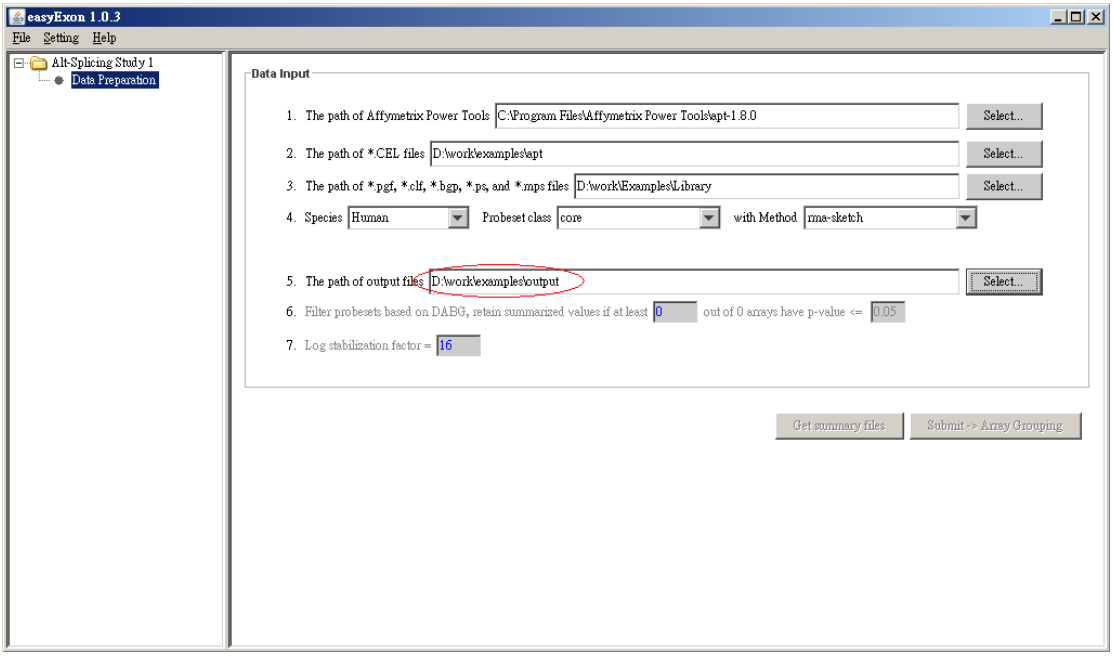


Method selection (rma-sketch or plier-gcbg-sketch)



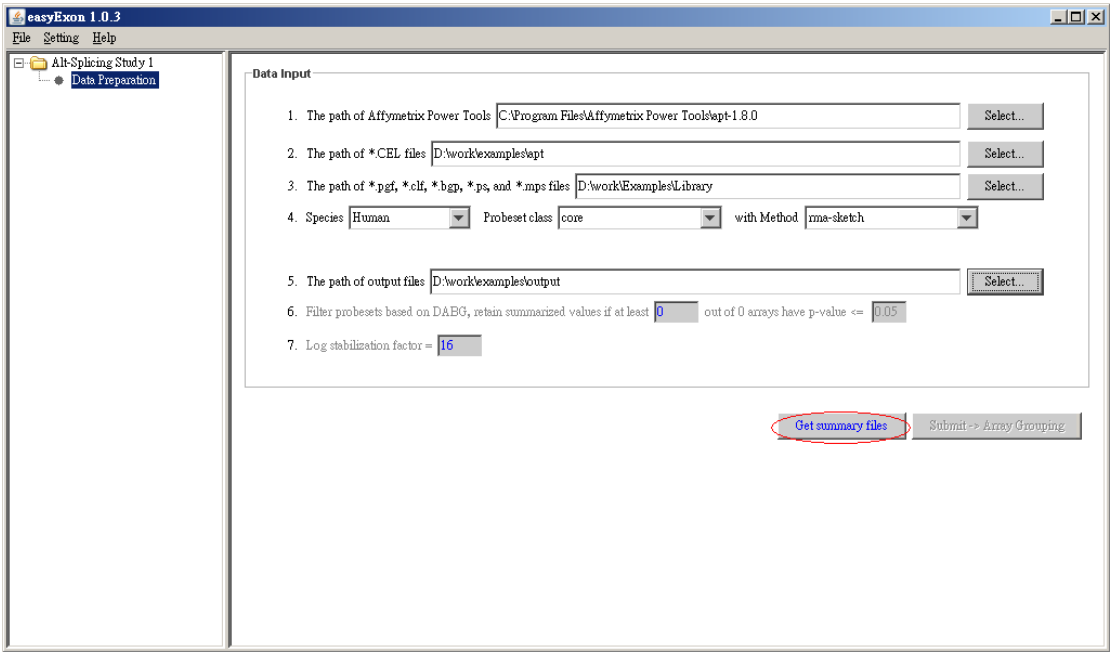
(v) Path of output files

Select the output directory for summary files



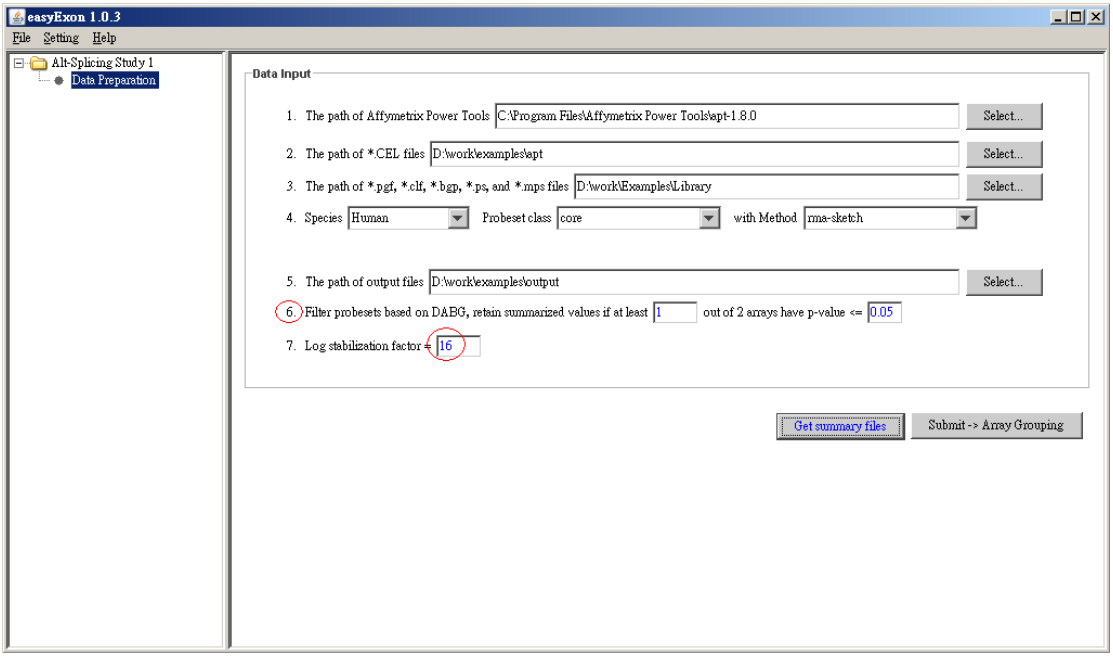
Get summary files!!!

Note: It may take few minutes to several hours, please have patience.

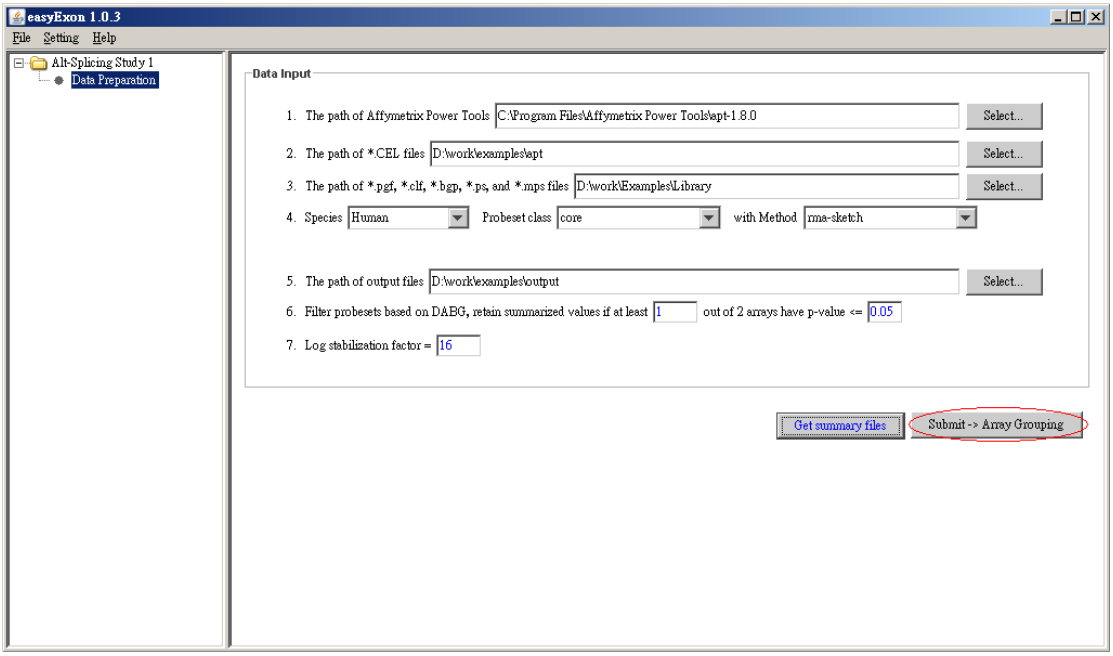


Filter probesets based on DABG, and set the log stabilization factor.

When the summary file is generated by APT, user can filter probesets based on DABG *p*-value. User can also set the log stabilization factor value. The summarized signal will add the factor before the log transformation. The default value is 16.



Submit to Step 2!!!



Step 1.b: Load *.txt summary files

Select the Exon-level summary files.

Select the data format (Expression Console or Partek) of Exon-level summary files, and set the path of file.

Data format:

(i) Expression Console:

The First row of the data contains “probeset_id” and CEL names. The first column contains probeset_ids, and remaining columns are intensity of each CEL with corresponding probeset_id.

E.g.:

`probeset_id<Tab>01_KL1.CEL<Tab>02_KL3.CEL<Tab>03_KL5.CEL<Tab>…`

`2315252<Tab>4.31<Tab>3.28<Tab>3.72<Tab>…`

`2315253<Tab>2.46<Tab>1.59<Tab>2.20<Tab>…`

(ii) Partek:

The First row of the data contains “Filename” and probeset_ids. The first column contains CEL names, and remaining columns are intensity of each probeset_id with corresponding. CEL.

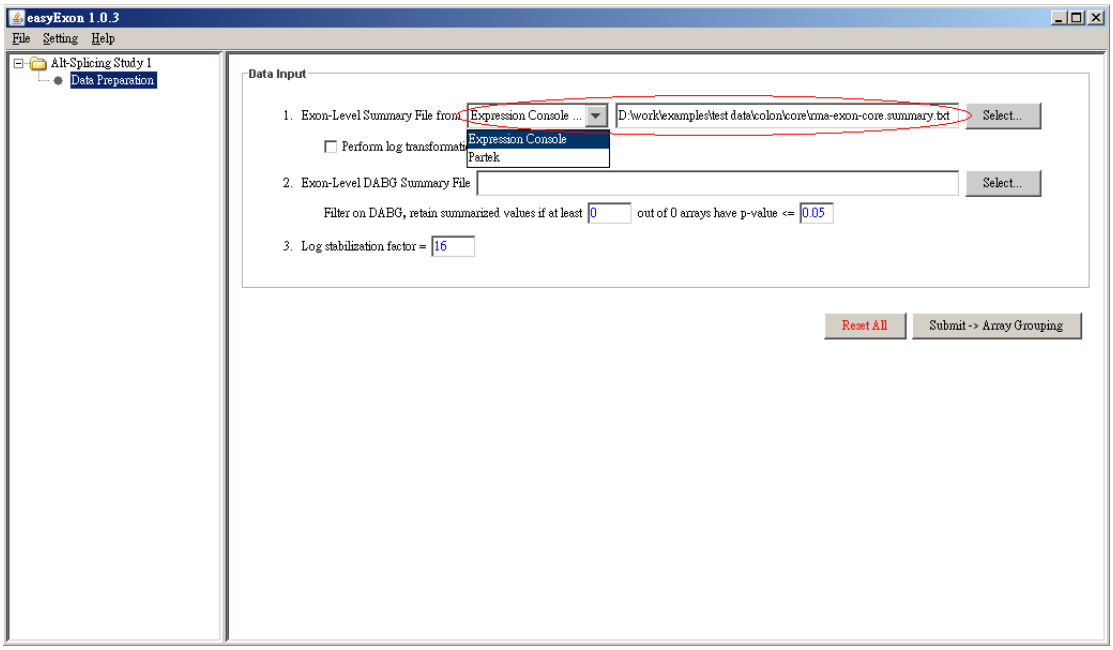
E.g.:

`Filename<Tab>…<Tab>2315252<Tab>2315253<Tab>…`

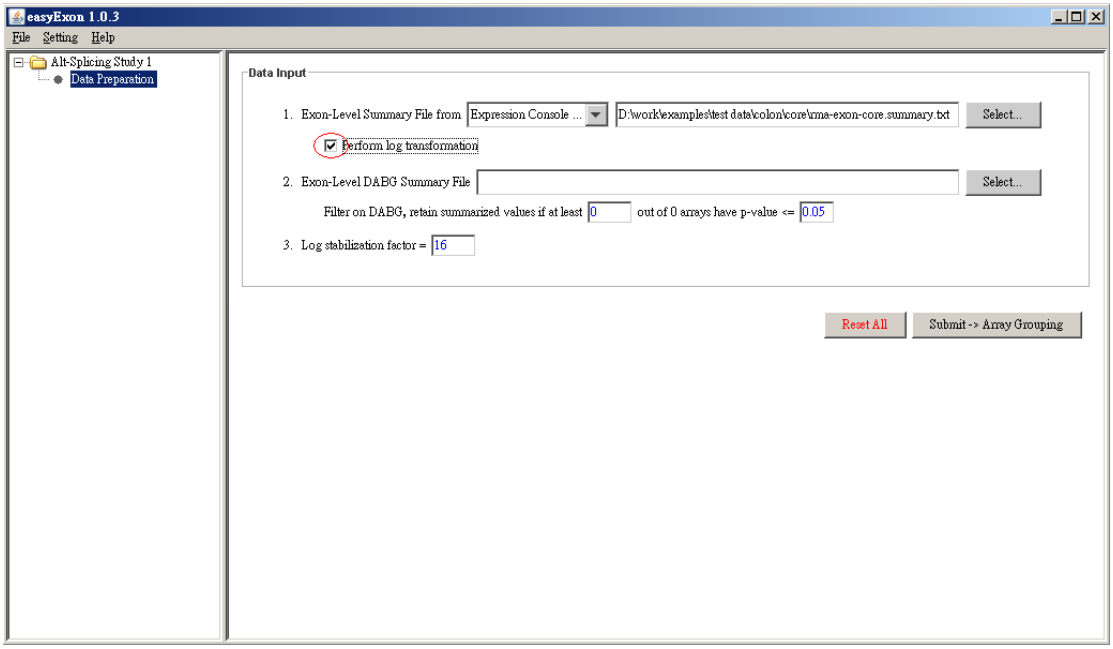
`01_KL1.CEL<Tab>…<Tab>4.31<Tab>2.46<Tab>…`

`02_KL3.CEL<Tab>…<Tab>3.28<Tab>1.59<Tab>…`

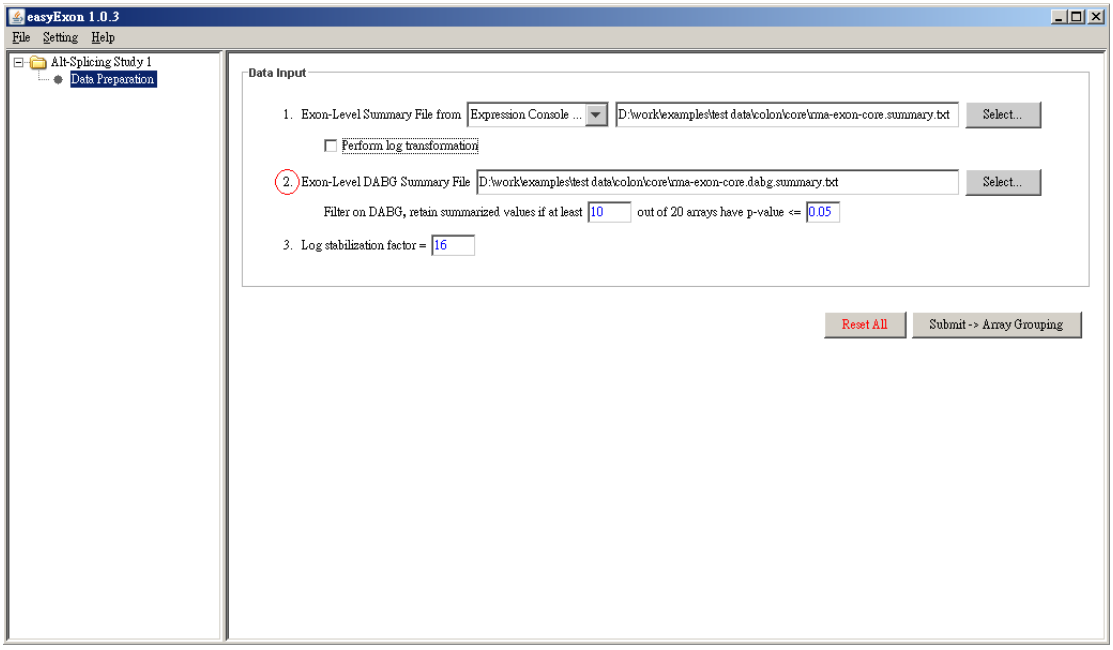
`03_KL5.CEL<Tab>…<Tab>3.72<Tab>2.20<Tab>…`



When user needs to perform log transformation on data, check the Perform log transformation.

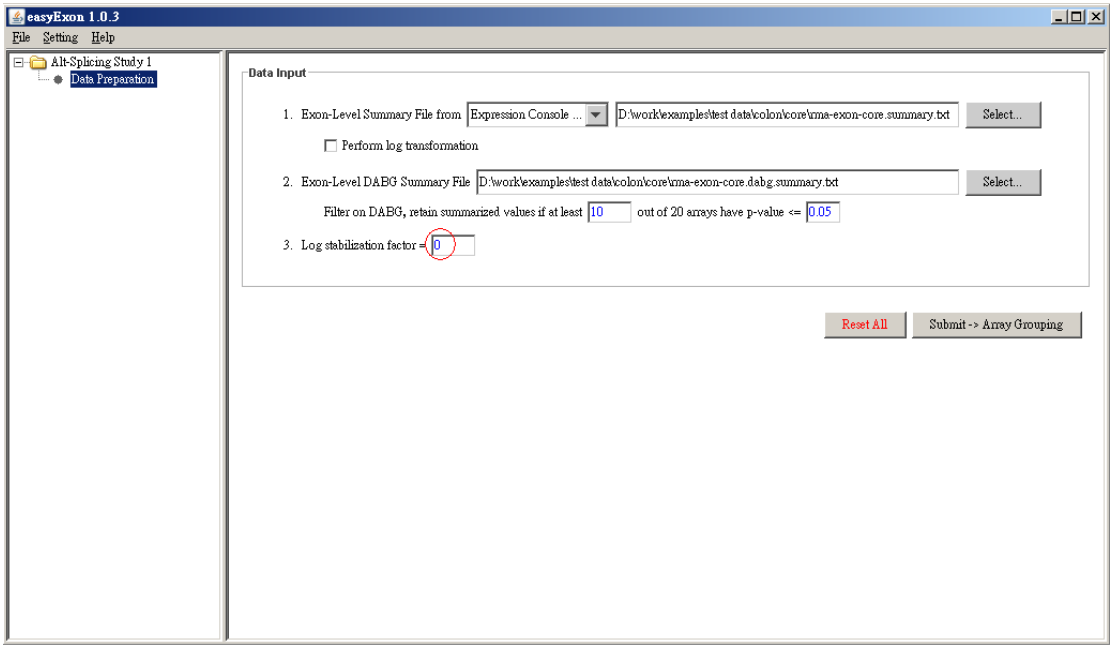


When user selects Expression Console, user needs to select the path of Exon-level DABG summary files and set probesets filter based on DABG *p*-value.

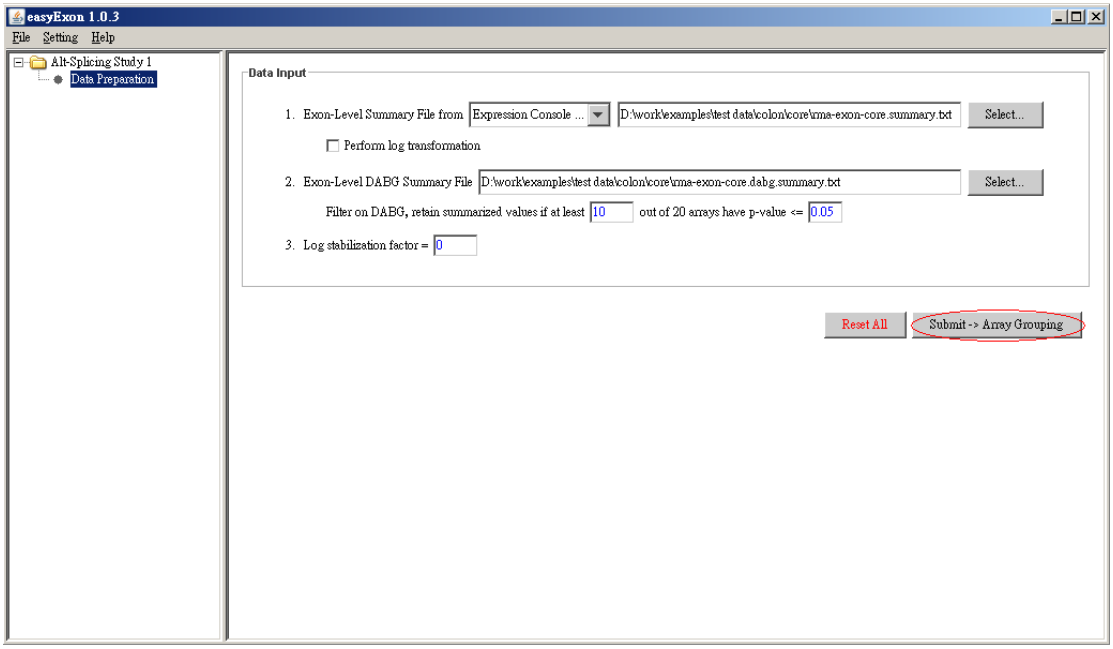


Set the log stabilization factor

When user selects Expression Console or Custom format, user needs to set the log stabilization factor. The summarized signal will add the factor before the log transformation. The default value is 16.



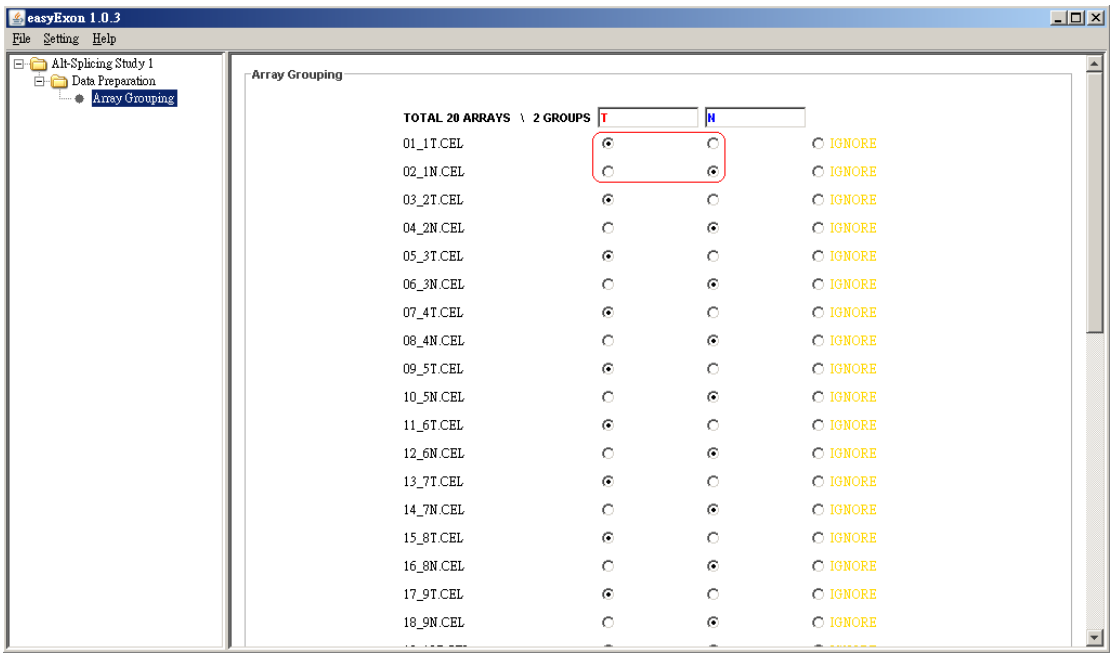
Submit to Step 2!!!



Step 2: Array Grouping and Data Processing

Step 2-1: Array Grouping

Array samples are assigned into two groups based on experimental design. The group name can be arbitrarily modified for easy recognition. For samples that don't want to be included in the latter analyses, they can be assigned into the "IGNORE" group.

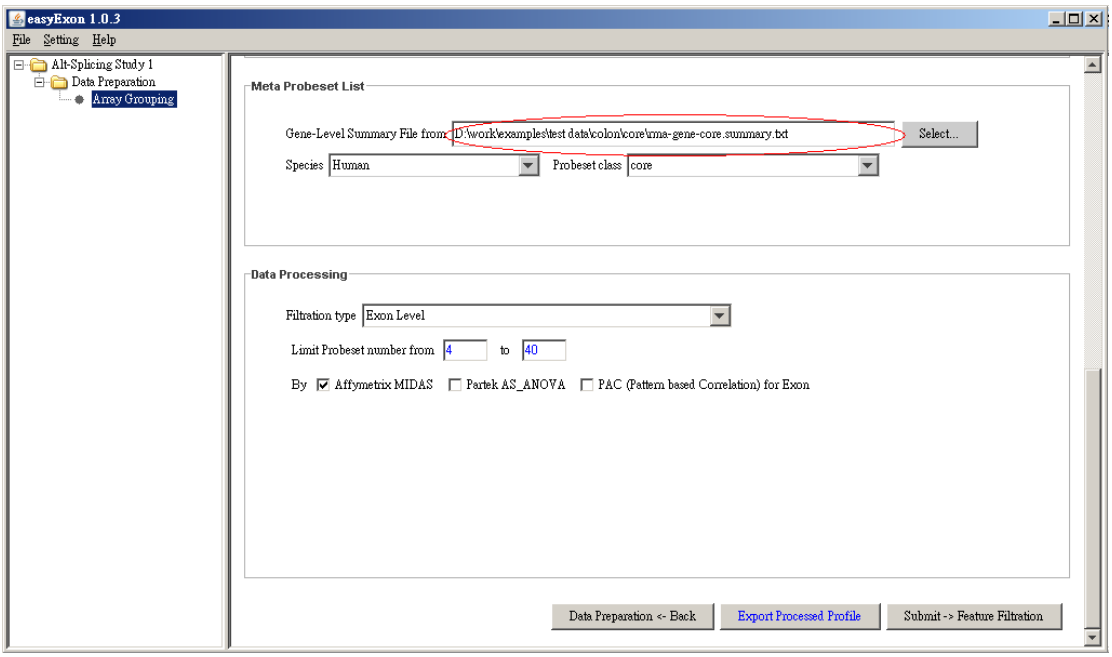


Step 2-2: Meta Probeset List Information

The information needed in this section is used for calculation of splicing index.

Step 2-2-1: Select Gene-level summary file

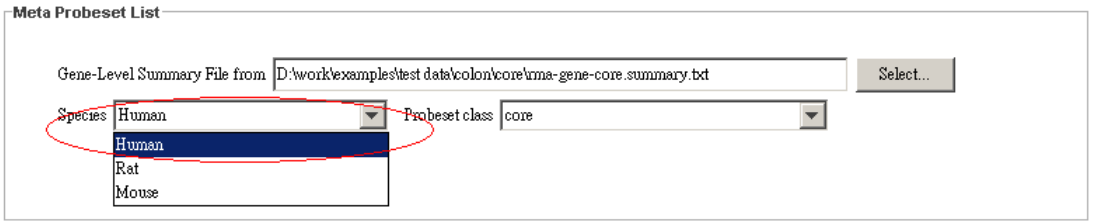
User who selects Expression Console at the Step 1.b needs to set the path of gene-level summary file. Otherwise, skip this step.



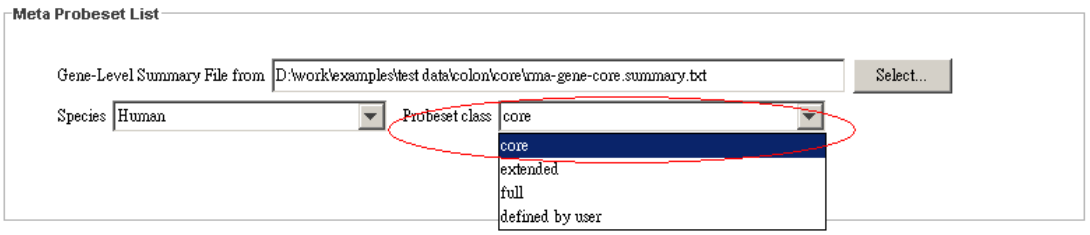
Step 2-2-2: Select meta probeset definitions

User who chooses NO at the Step 0 needs to select the Species and Probeset class for meta probeset definitions. Otherwise, skip this step.

Species selection (Human, Rat, or Mouse)



Probeset class selection (core, full, extended, or defined by user)



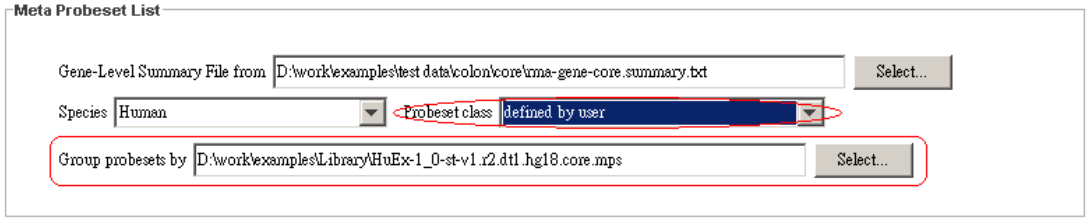
When user selects “defined by user,” user needs to select the meta-probesets file location.

Meta-probesets file format:

The First row of the data contains “probeset_id”, “transcript_cluster_id”, and “probeset_list”. The first column contains probeset_ids; the second column is transcript_cluster_id. The third column is probeset_list with corresponding transcript_cluster_id.

E.g.:

```
probeset_id<Tab>transcript_cluster_id<Tab>probeset_list<Tab>…
3096575<Tab>3096575<Tab>3096630 3096631 3096632<Tab>…
2949118<Tab>2949118<Tab>2949119 2949120 2949121 2949130<Tab>…
```



Step 2-3: Data Processing

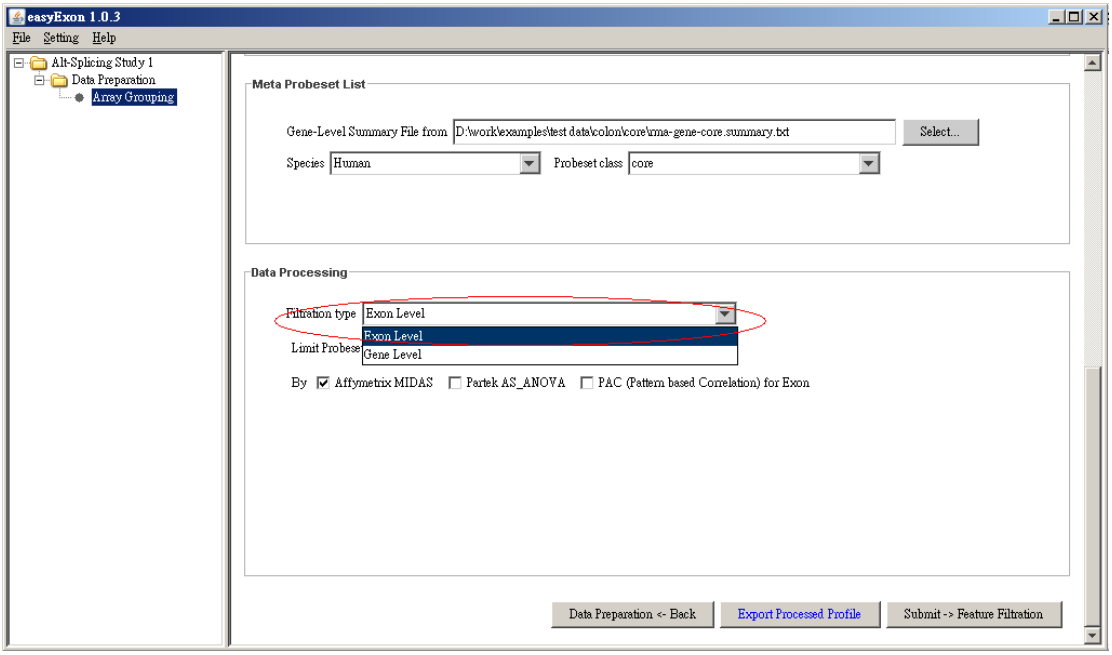
There are two types of differential expression filtrations in easyExon: Exon level and Gene level.

(i) Exon Level

Choose Exon Level and follow the Step 2-3.a.

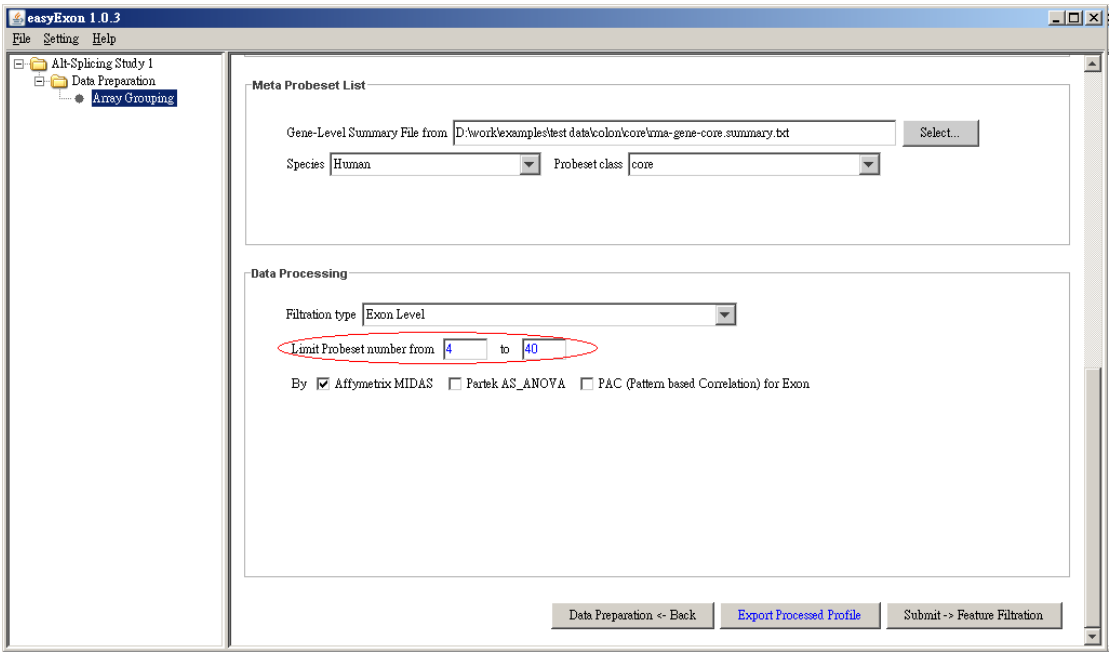
(ii) Gene Level

Choose Gene Level and follow the Step 2-3.b.



Step 2-3.a: Exon Level Statistical Filtration

Before performing the statistical flirtation, user sets the limitation number of probesets in a transcript cluster. For the transcript cluster with probesets greater than the setting number, only the setting number of probesets will be included in the latter analysis.



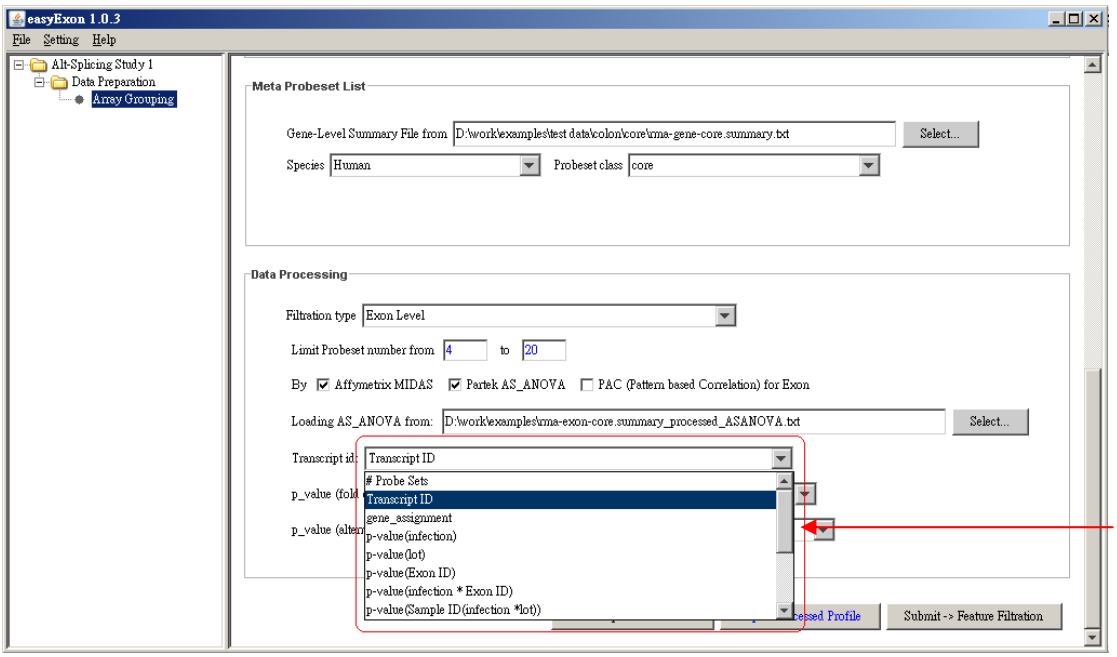
There are three statistical methods for Exon level filtration in easyExon: Affymetrix MIDAS (Microarray Detection of Alternative Splicing), Partek AS ANOVA, and PAC (Pattern-Based Correlation) for Exon. User can use multiple selections to choose the more than one method at the same time.

(i) Affymetrix MIDAS

Default selection for statistical filtration. No other information is needed.

(ii) Partek AS ANOVA

User needs to set the path of AS ANOVA file, and sets the attributes of Transcript id and *p*-values from attribute list, which is the first row of the data, and each column is an attribute. Attribute list differs from AS ANOVA files.

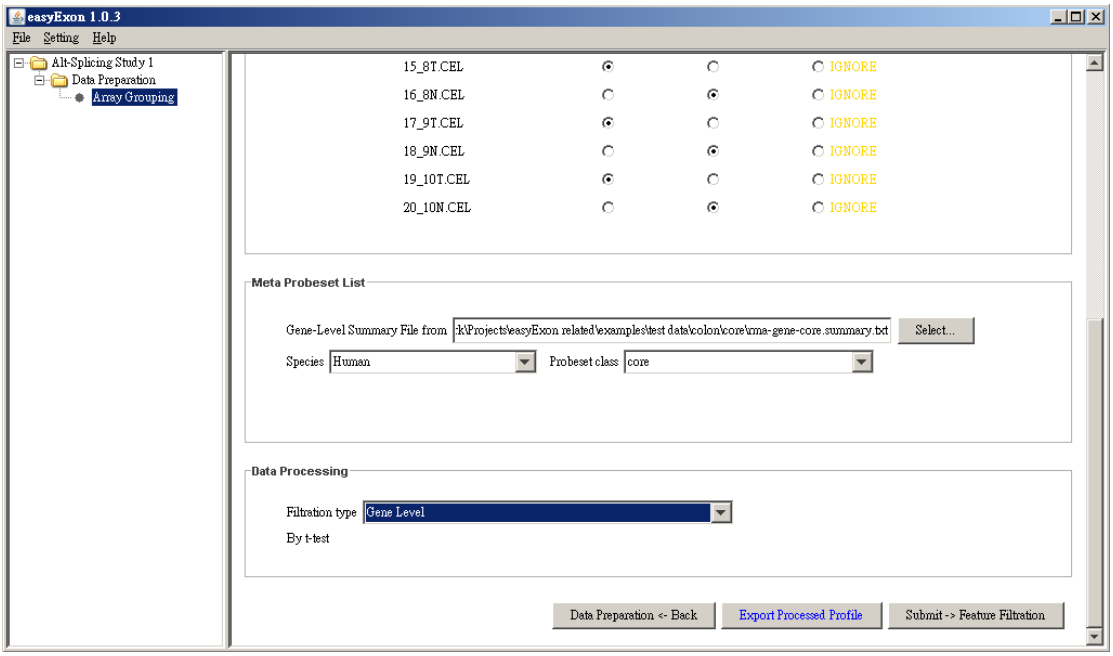


(iii) PAC (Pattern based correlation) for Exon
No other information is needed.

Last, go to the Step 2-4.

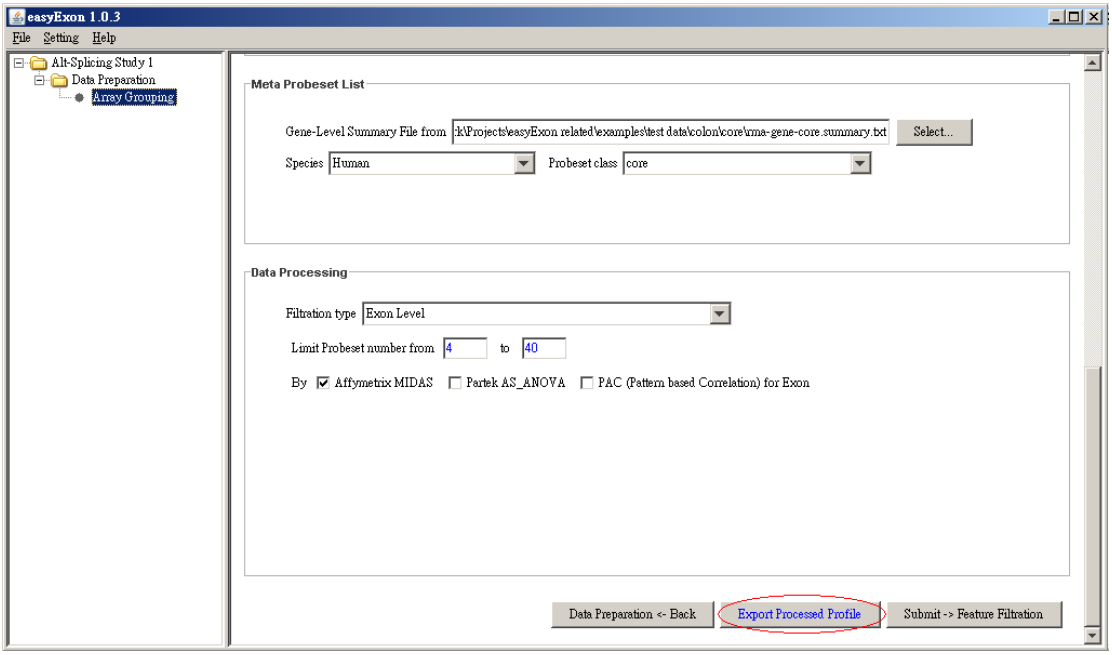
Step 2-3.b: Gene Level Statistical Filtration

easyExon will perform t-test filtration on gene level in this step. No other information is needed.

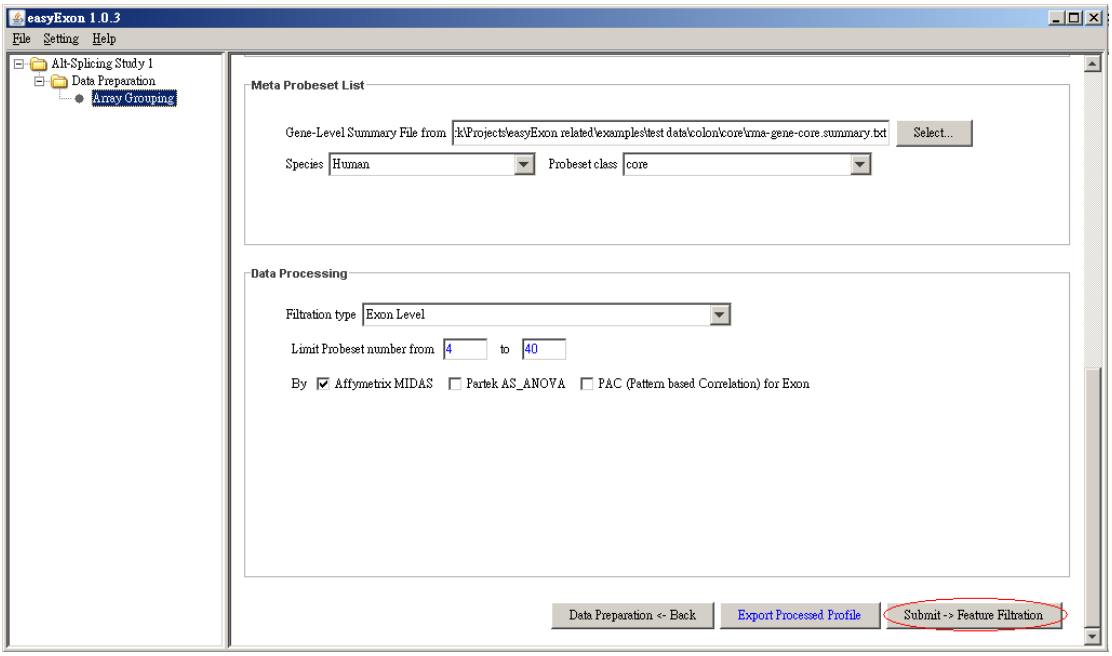


Step 2-4: Export Processed Profile (Optional)

Export the processed file with log stabilization.



Submit to Step 3!!!

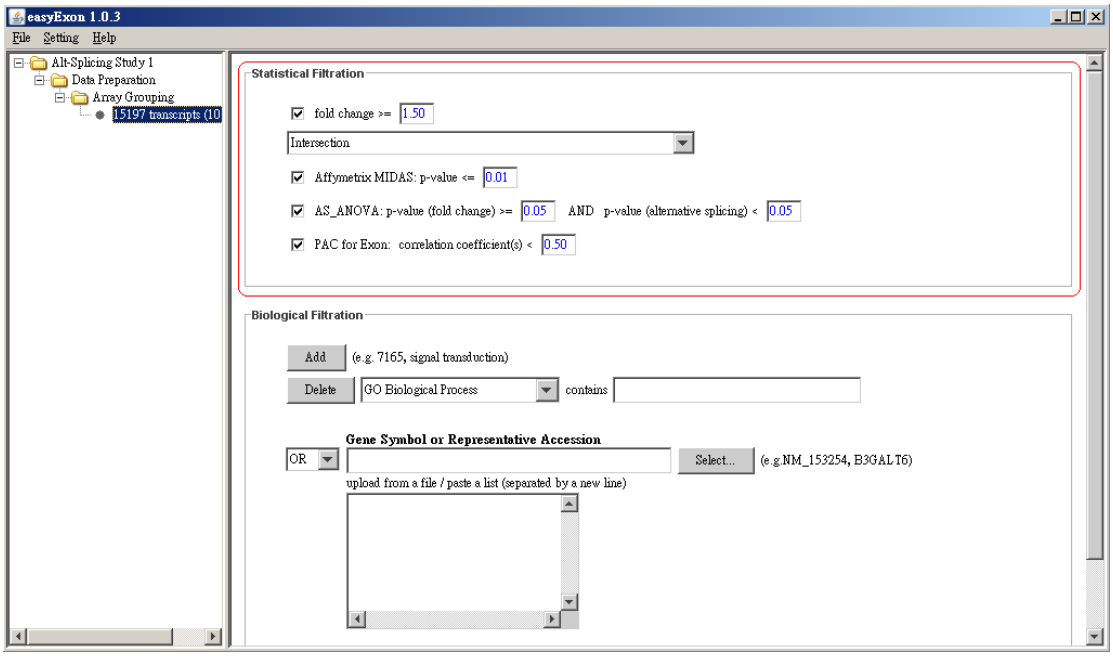


Step 3: Feature Filtration

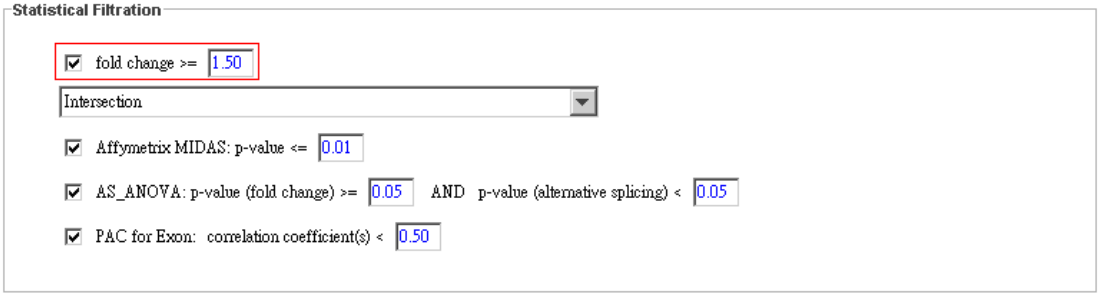
When user chooses Exon Level at the Step 2-3.a, please follow the Step 3-1.a. Otherwise, please follow the Step3-1.b.

Step 3-1.a: Statistical Filtration (Exon Level)

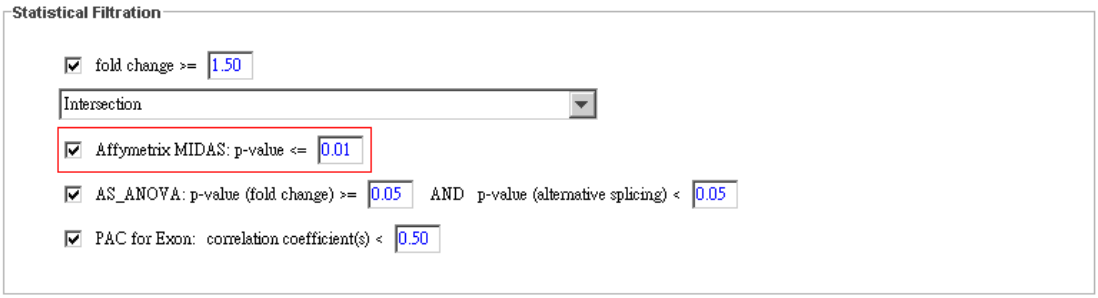
When user chooses method(s) for statistical filtration at the Step 2-3.a, user needs to set the criteria for probeset filtration.



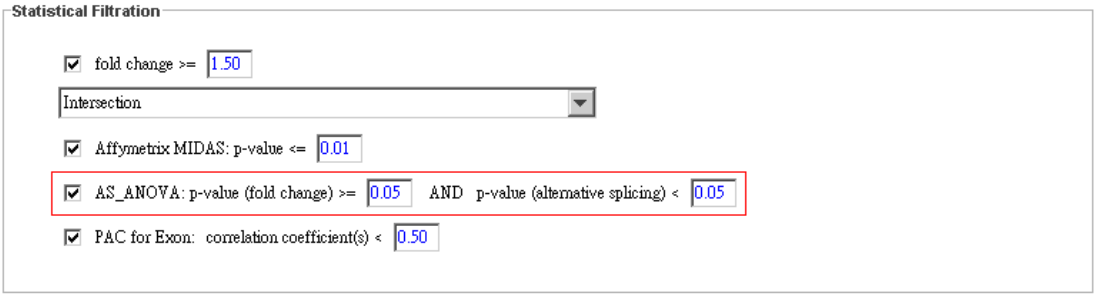
First, user needs to set the fold change.



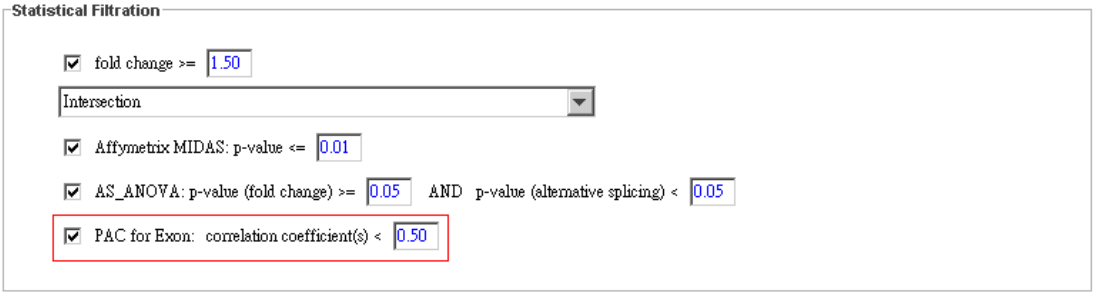
When user chooses Affymetrix MIDAS at the Step 2-3.a, user needs to set the *p*-value (calculated by MIDAS).



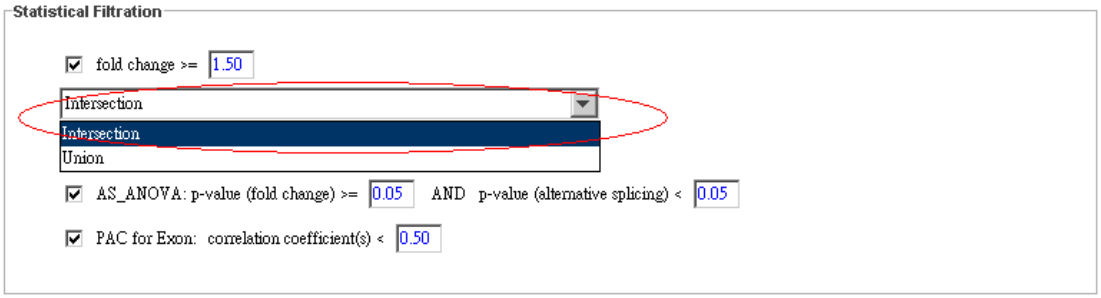
When user chooses AS_ANOVA at the Step 2-3.a, user needs to set the *p*-value for fold change and *p*-value for alternative splicing. Otherwise, skip this step.



When user chooses PAC for Exon at the Step 2-3.a, user needs to set the dissimilarity value of correlation coefficients (the default value is less than 0.5).



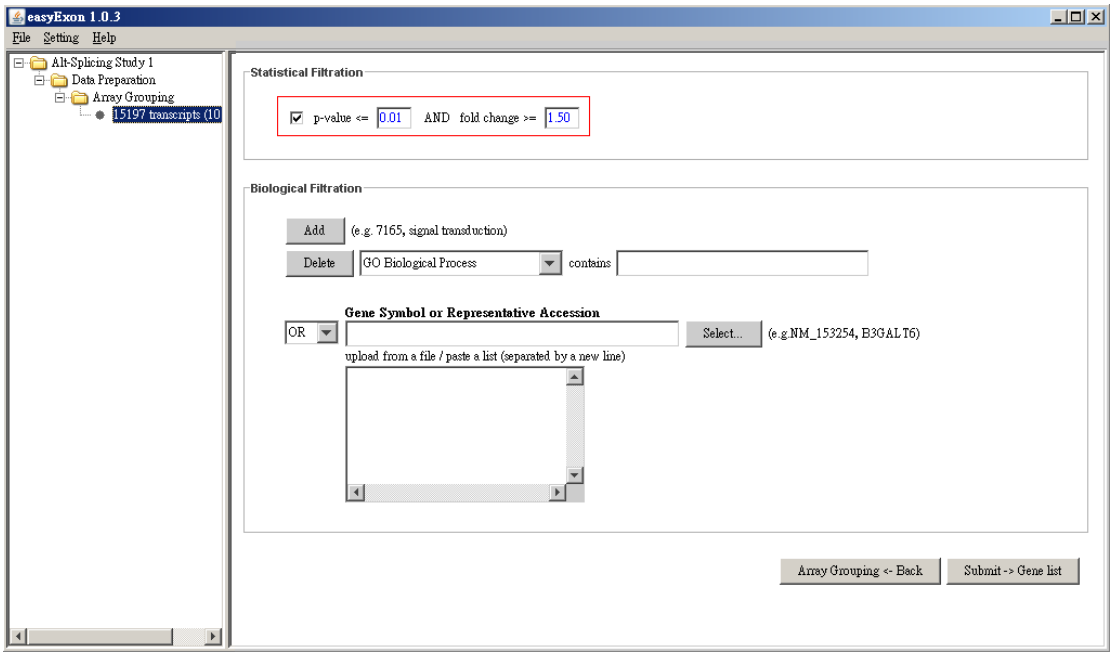
When user chooses more than one probeset filtration methods, user can select different ways to combine the filtered results.



User can jump to the Step 3-2.

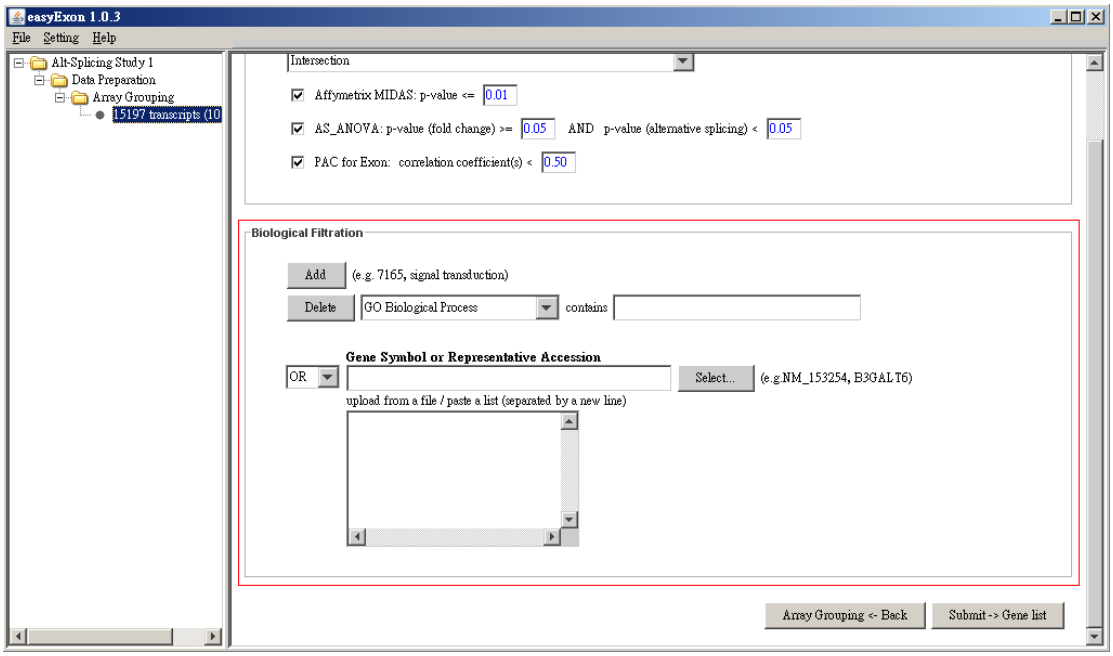
Step 3-1.b: Statistical Filtration (Gene Level)

User sets the significant level of *p*-value (calculated by *t*-test) and the minimum value of fold change.



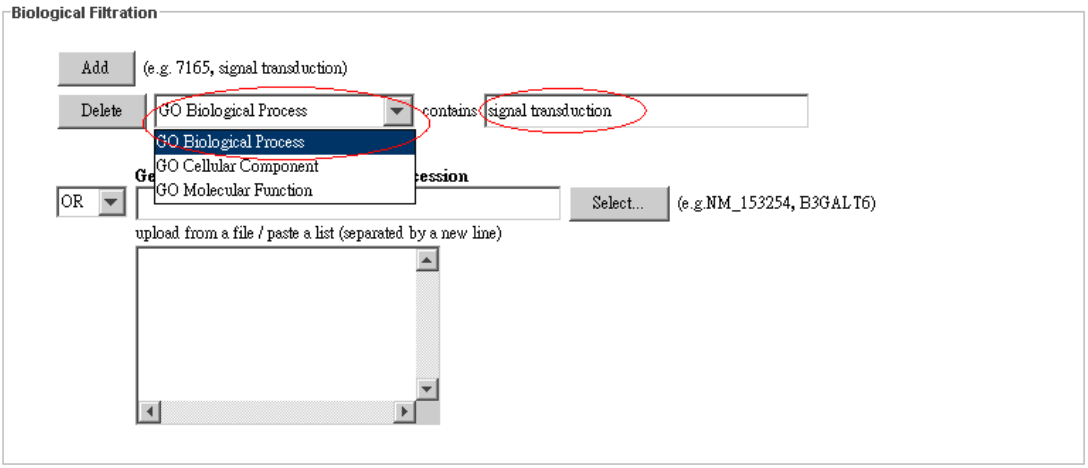
Step 3-2: Biological Filtration

Transcript cluster filtration tool aims to help user narrow down the number of transcripts based on (i) GO term and (ii) gene name or accession number. User can add queries for transcript cluster filtration.

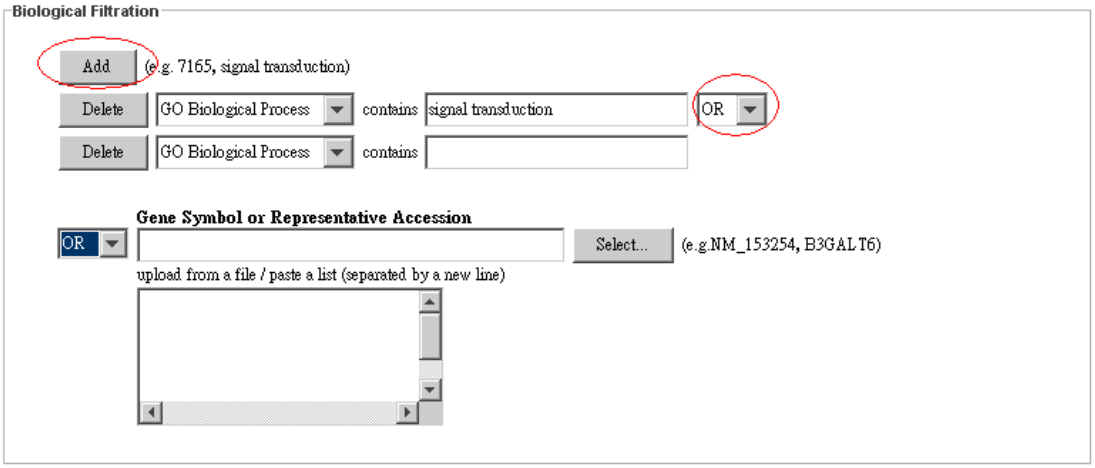


(i) Query by GO term

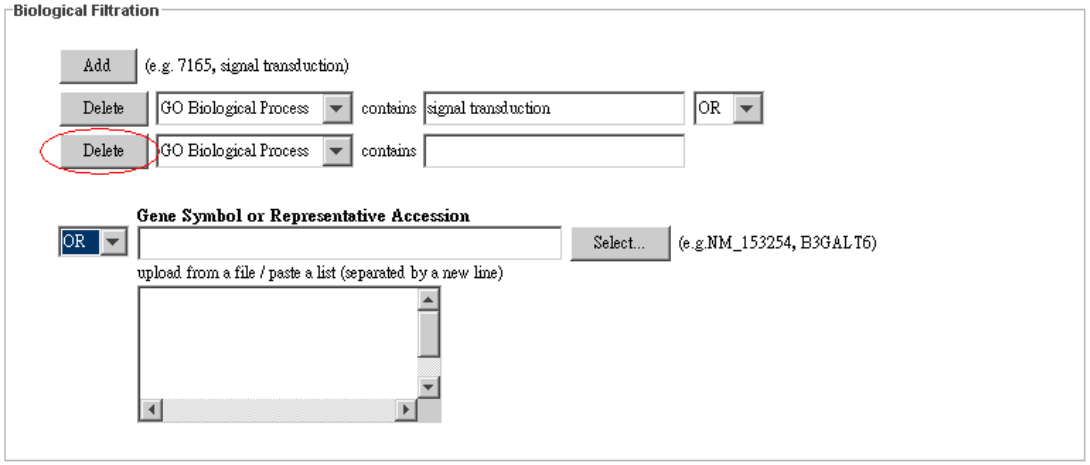
User can choose the GO term (e.g. Biological Process), and type the key word in the text field (e.g. signal transduction).



When there are more than one queries, user may click the “Add” button to add a new GO query and choose “AND” or “OR” for the query. Ten queries are allowed.

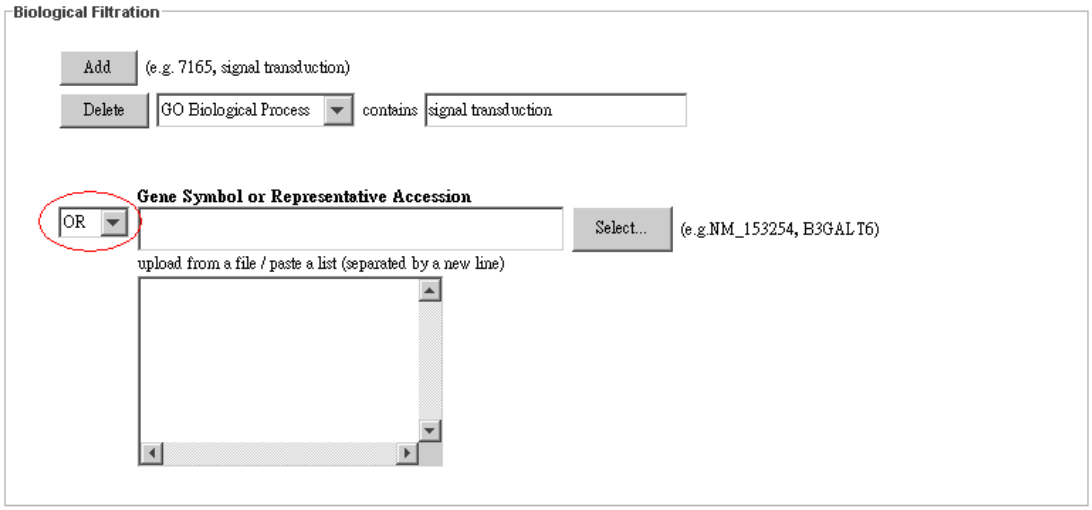


When user wants to delete one query, user may click the “Delete” button in front of the query.

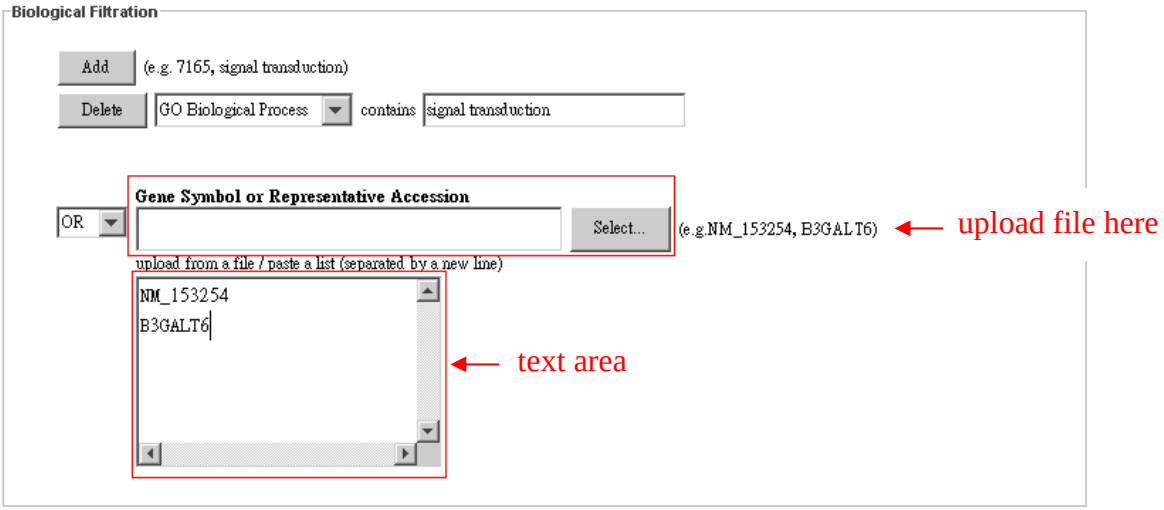


(ii) User can query Gene Symbols or Accession numbers of interest by uploading a file or using text area.

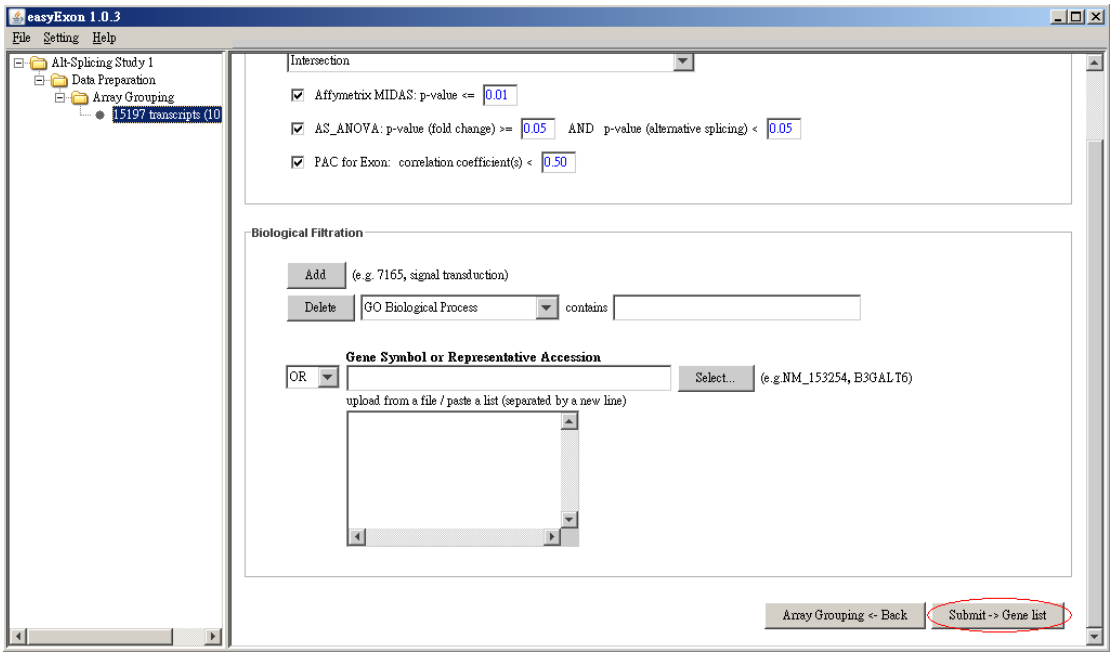
When user sets the GO query above, user may choose “AND” or “OR” to combine with Gene Symbol’s queries.



Type the key words in the text file or the text area below. Each key word is separated by a new line.



Submit to Step 4!!!

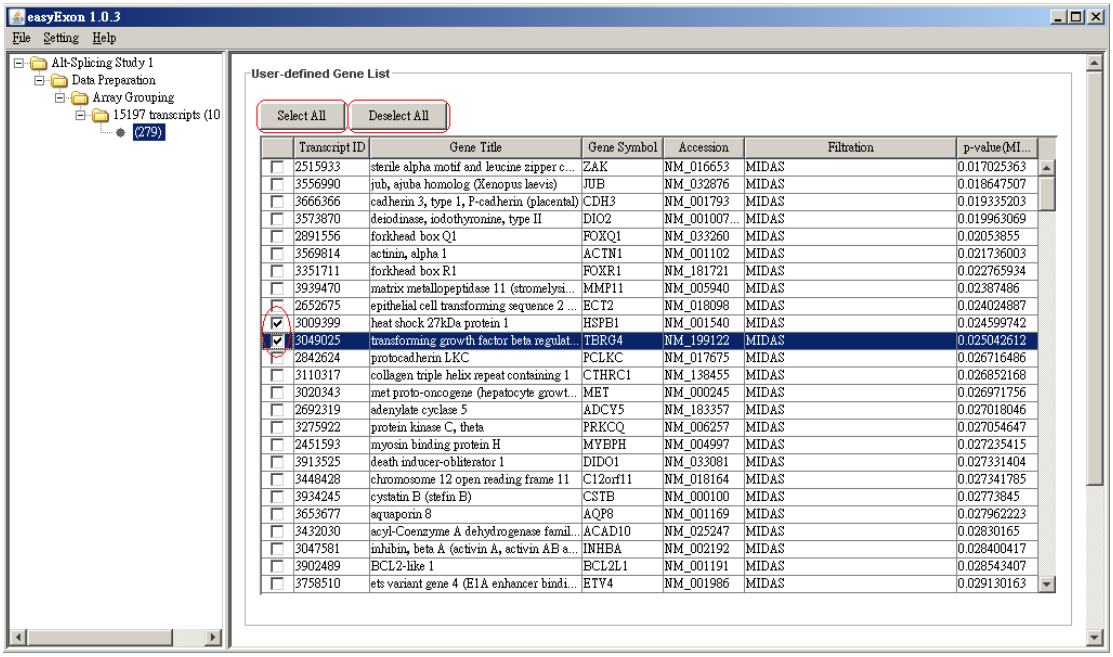


Step 4: Gene List

Step 4-1: User-defined Gene List

Further select genes of interest. Genes in the narrowed down list will be displayed with log scale intensity in the Step 5. The corresponding number of selected genes will be shown in highlighted in the directory tree of the left panel.

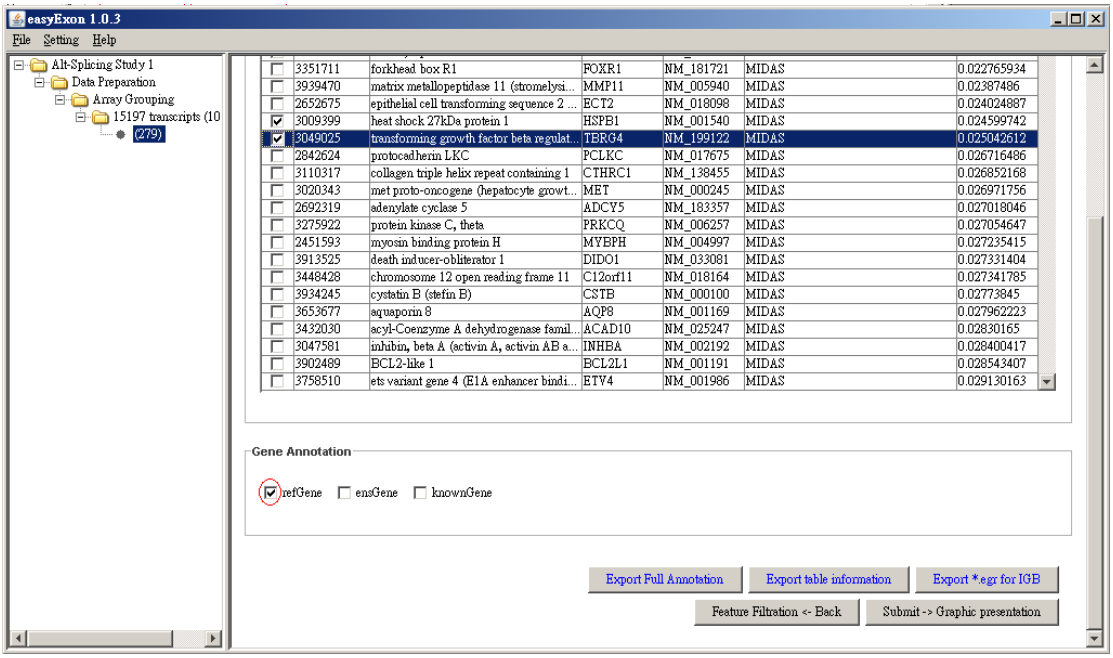
User can use the “Select All” or “Deselect All” button to further selection.



Step 4-2: Gene Annotation

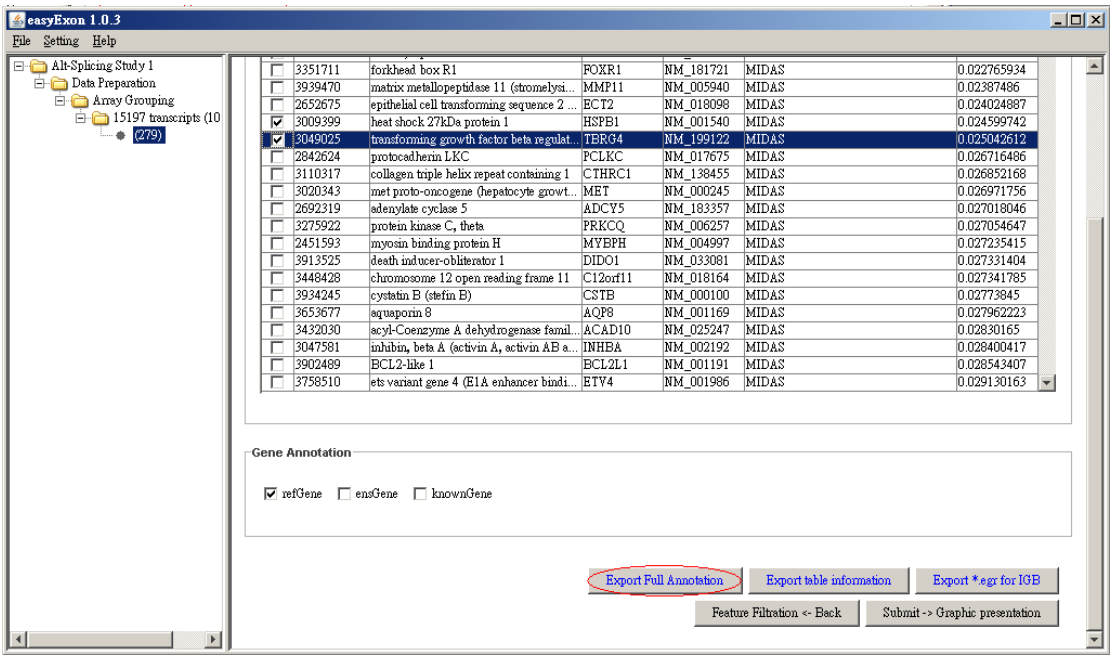
Note: User who chooses “defined by user” for probeset class at the Step 1.a or the Step 2-2-2 needs to skip this step. In addition, user who chooses “Gene Level” at the Step 2-3 skips this step, too.

User selects the gene annotation definition(s) for graphic presentation.



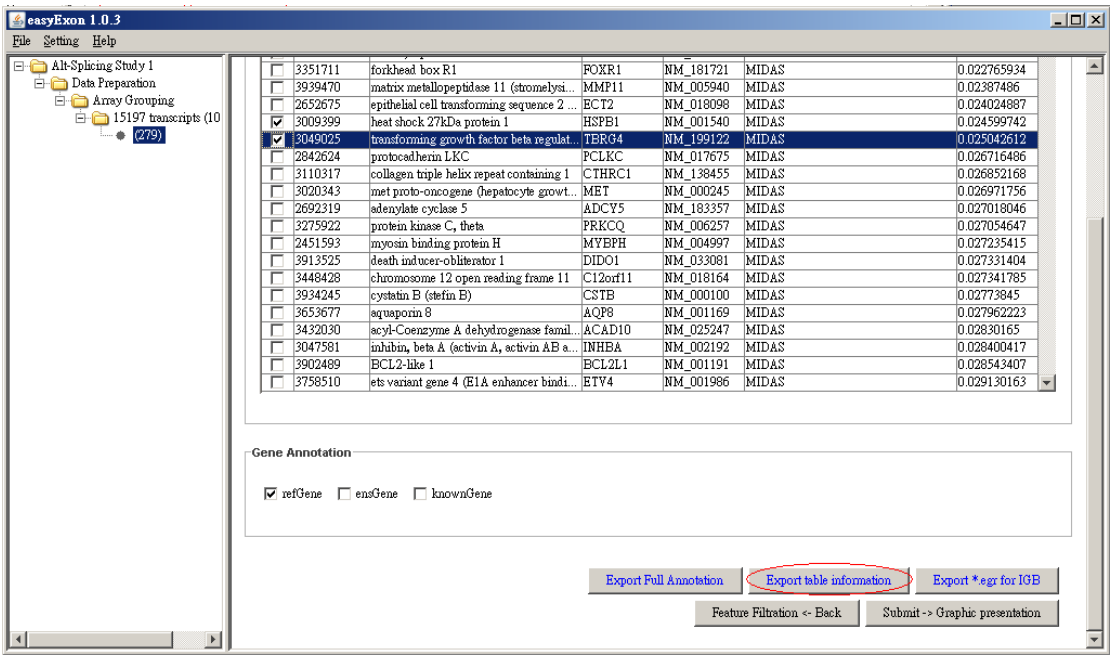
Step 4-3: Export Full Annotation (Optional)

User may export full annotation by click the button “Export Full Annotation.”



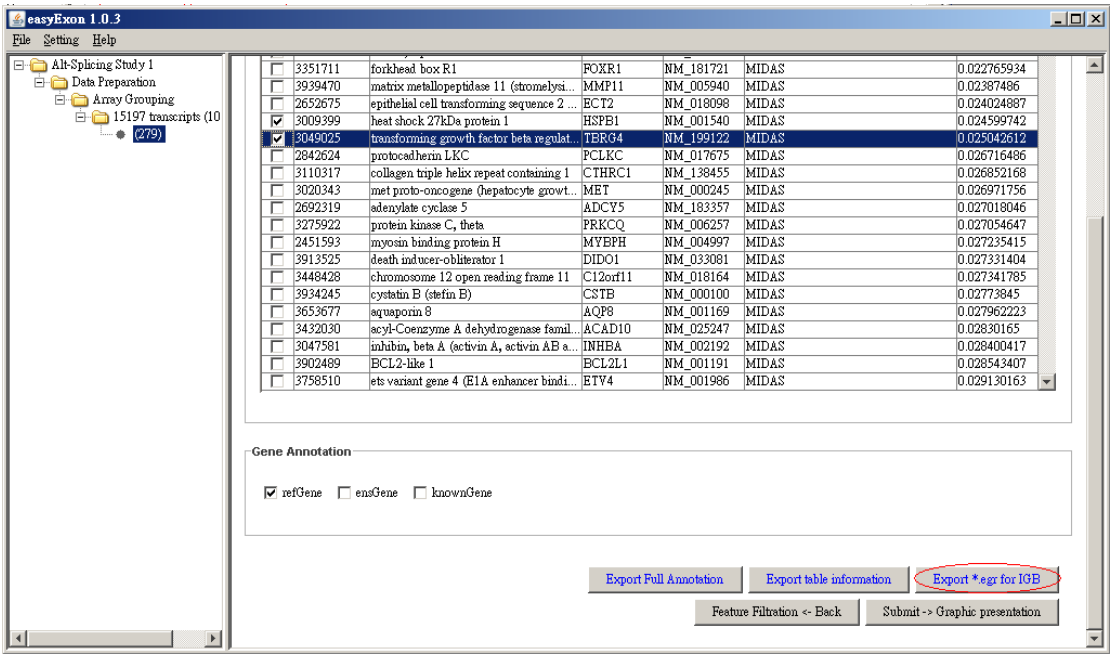
Step 4-4: Export Table Information (Optional)

User may export table information by click the button “Export table information.”

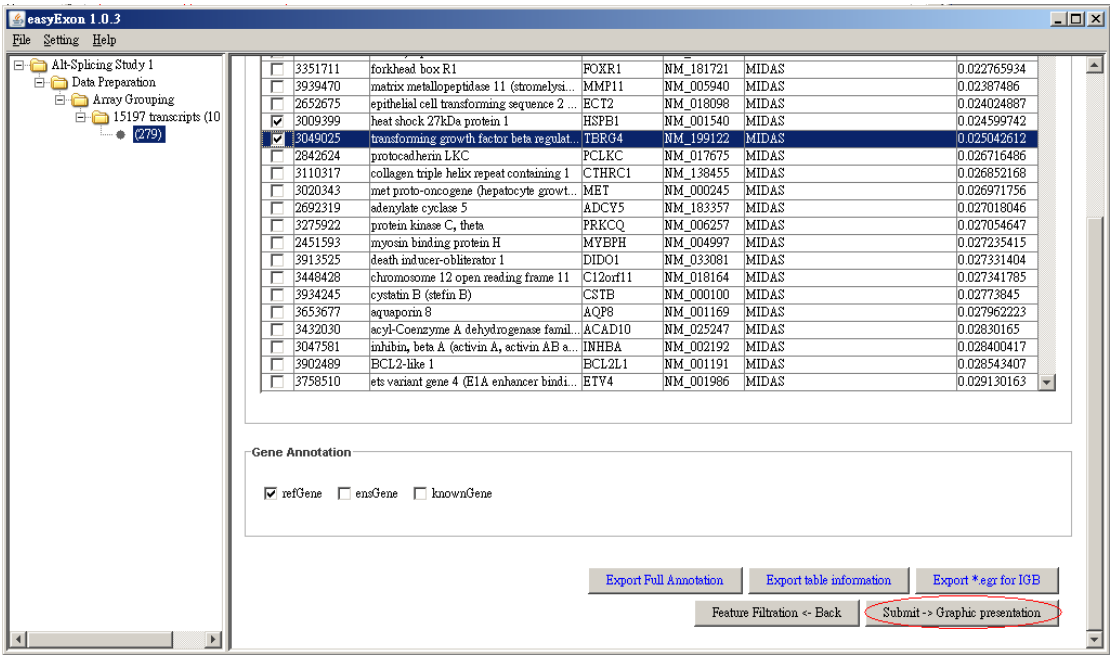


Step 4-5: Export *.egr for IGB (Optional)

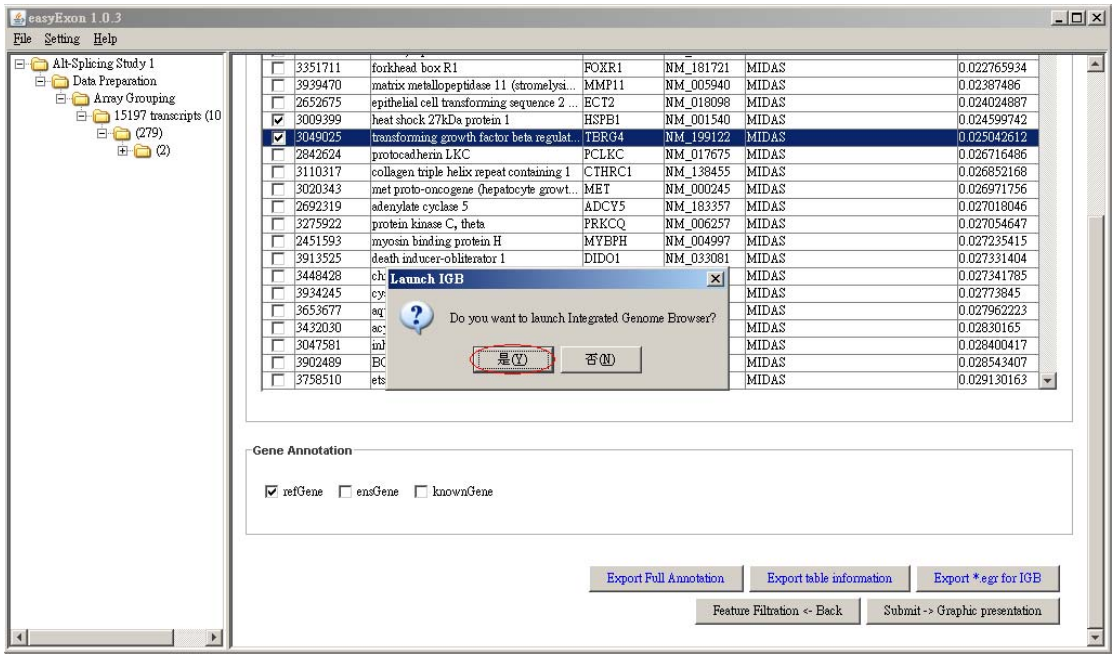
User may export a *.egr file for Integrated Genome Browser (IGB) by clicking the button “Export *.egr for IGB.”



Submit to Step 5!!!



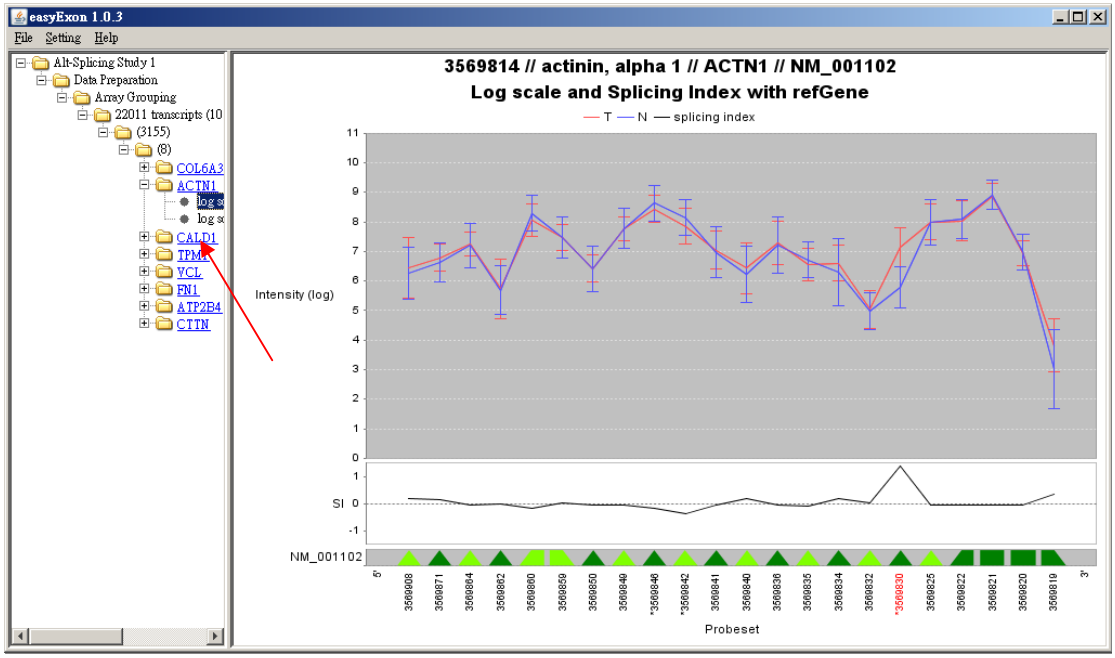
When user needs to launch IGB for browsing the transcript information, user needs to choose “YES” to launch the IGB.



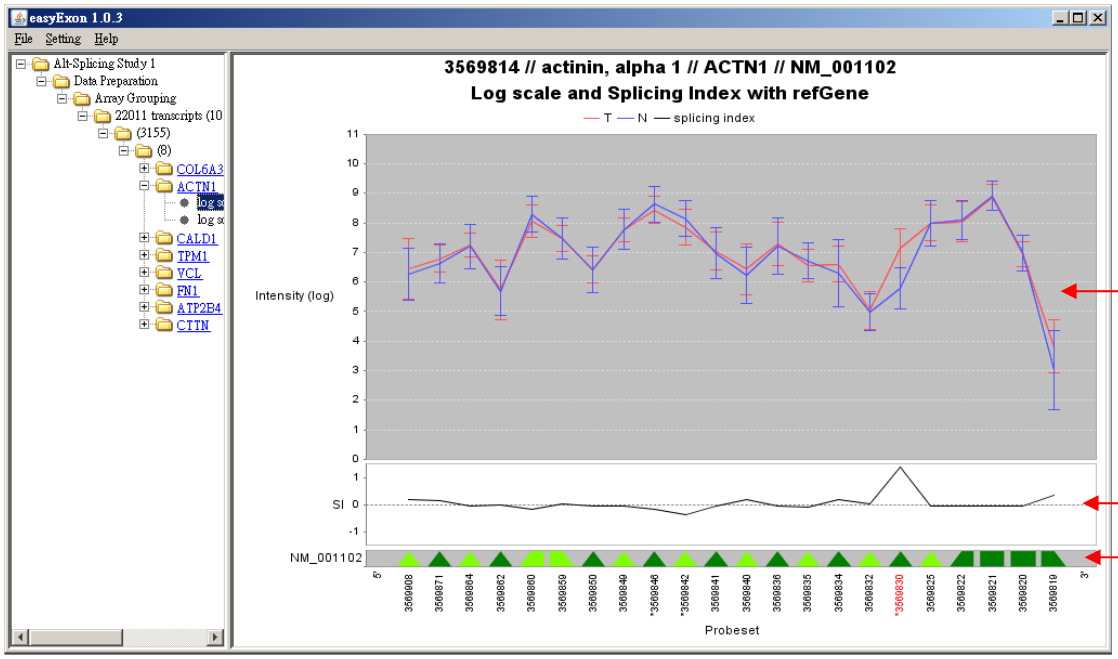
Step 5: Graphic Presentation

Step 5.a: Graphic Presentation for Exon Level

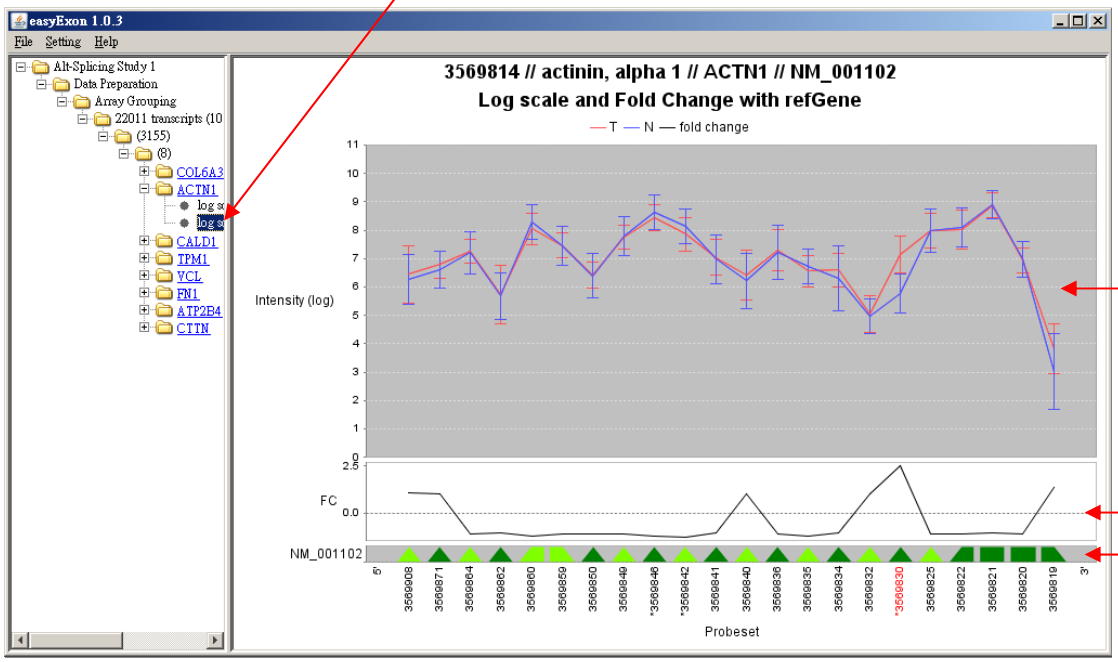
When user launches IGB, user may click the hyper link in the left panel to browse the transcript in IGB.



When user clicks the gene of interest in the left panel, a graphic representation for exons in the corresponding transcript cluster will be displayed in the right panel. The x-axial is the probeset id for each exon, and y-axial represents log scale intensity and splicing index. Neighboring probesets with the same color in gene annotation means that those probesets are in the same exon.



User may select log scale and fold change on the tree of left panel, intensity of log scale and fold change will be displayed on the right panel.



User can modify the Filtration (Step 3), and gene signature and gene annotation (Step 4) repeatedly.

Note: Labels for probeset id of x-axial.

(i) Probeset id labeled in red: fold change is greater than the threshold set in Step 3-1.a.

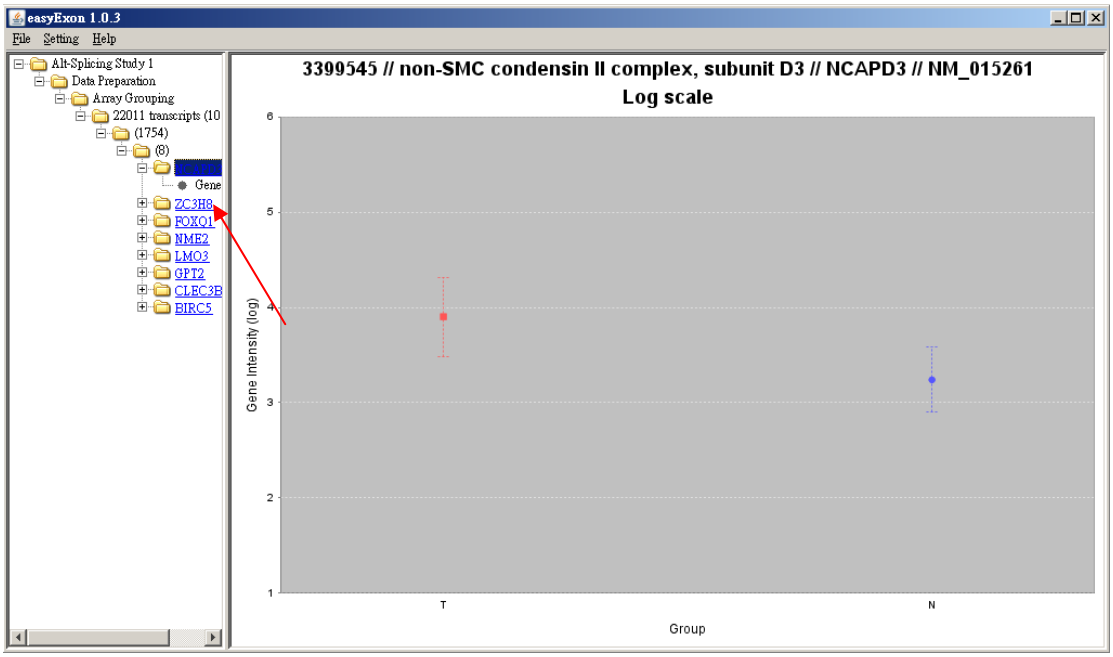
(ii) Probeset id labeled in light gray: *p*-value from DABG file is greater than the threshold set in Step 1.b.

(iii) Probeset id labeled with "*": *p*-value computed by MIDAS is smaller than the threshold set in Step 3-1.a.

(iv) Probeset id labeled with "PAC": Patter based correlation is smaller than the threshold set in Step 3-1.a.

Step 5.b: Graphic Presentation for Gene Level

When user launches the IGB, user may click the hyper link to browse the transcript in the IGB.



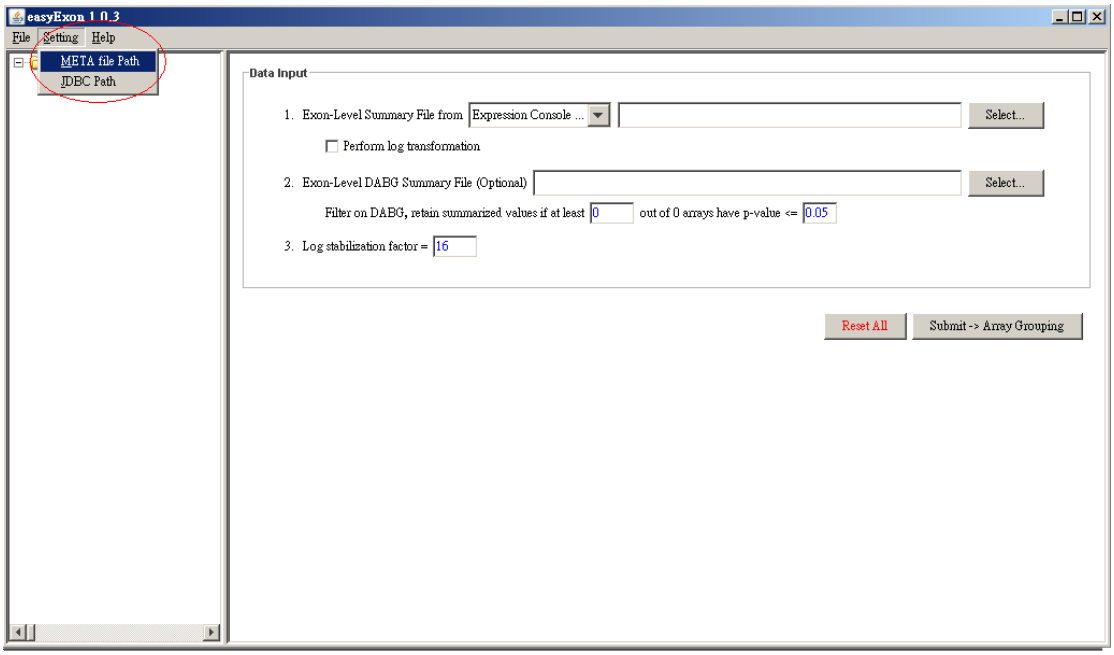
User can modify the Filtration (Step 3), and gene signature and gene annotation (Step 4) repeatedly.

2. Environment Setting

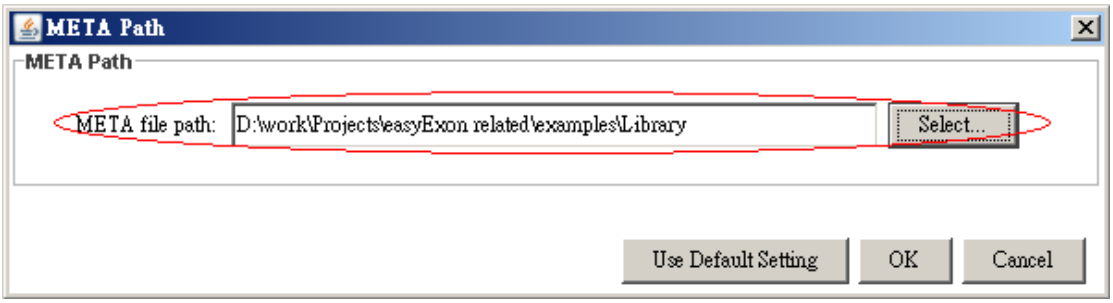
User may download our database and meta probeset files from our homepage (http://microarray.ym.edu.tw:8080/easyexon/index.jsp?mode=support) and set the path of those downloading files and database.

2-1: Setting the Path of Meta Probeset Files

User may set the path of meta probeset files by clicking “Setting” on Menu Bar and select “META file Path.”

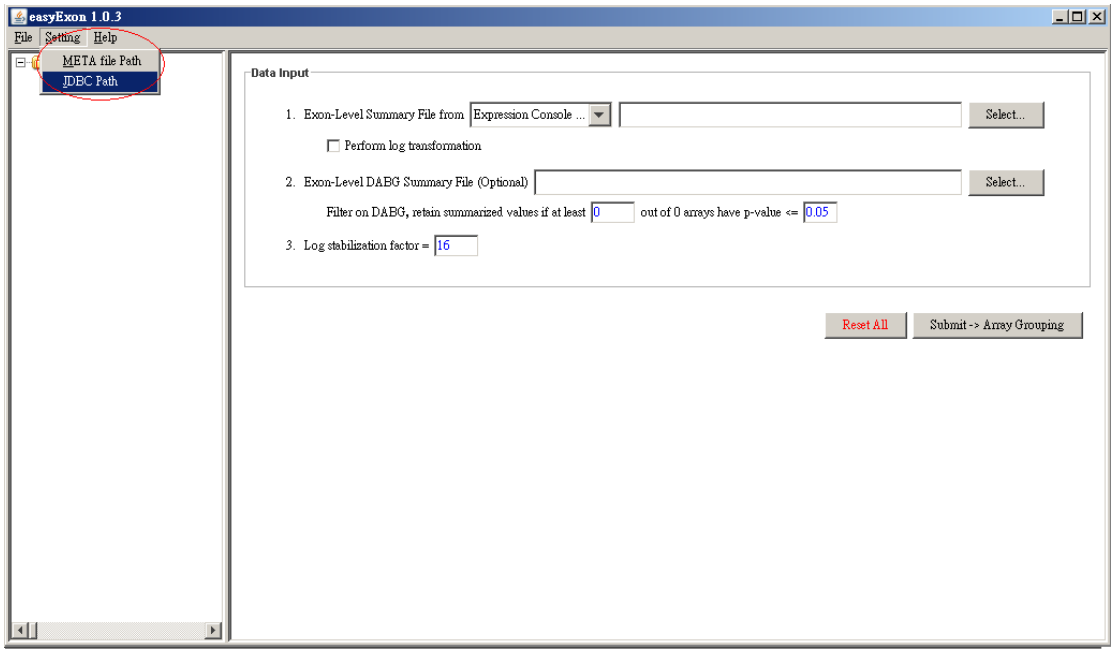


Select the path of meta probeset files directory which includes *.mps files.

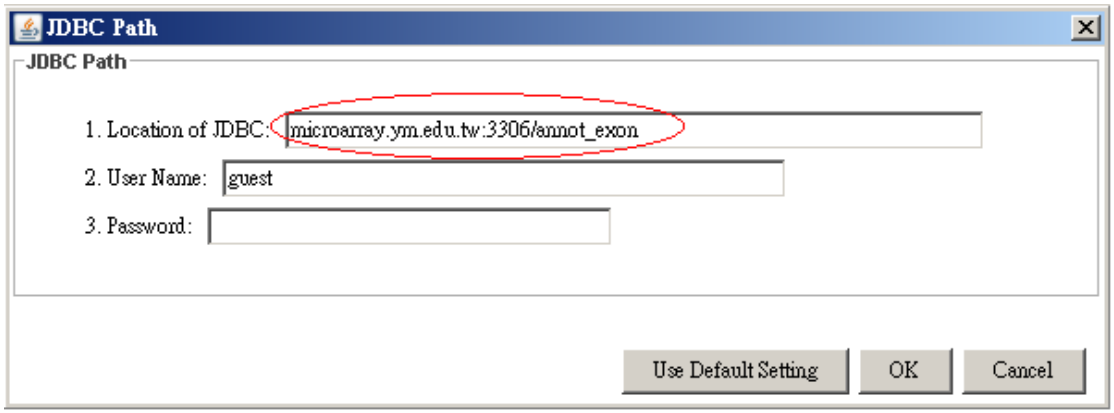


2-2: Setting the Path of Database

User may set the path of database by clicking “Setting” on Menu Bar and select “JDBC Path.”



Type the location of database which includes IP address (or domain name), port number, and the name of database. (**E.g.:** microarray.ym.edu.tw:3306/annot_exon)



Type the user name and password.