MICROARRAY *Я* US USER GUIDE

VERSION 2 (MARCH 2013)

YILIN DAI[1], LING GUO[1], MENG LI[2], AND YIBU CHEN[2]

1. Department of Mathematical Sciences, Michigan Technological University. Houghton, MI 49931.

2. Bioinformatics Service Program, Norris Medical Library, University of Southern California. Los Angeles, CA 90089.

CONTENTS

CHAPTER 1. INTRODUCTION

Featuring a user-friendly graphic interface, Microarray Я US is an R-based program that integrates functions from a dozen or so most-widely used Bioconductor packages (Gentleman, Carey et al. 2004) to offer researchers a streamlined way to perform routine microarray expression data analysis without the need of learning R language (Development Core Team 2011).

Supporting major expression microarray chips from both Affymetrix (Affymetrix, Santa Clara, CA) and Illumina (Illumina Inc., San Diego, CA), Microarray Я US provides a complete workflow that covers the following tasks:

- ✓ Data import (supply by users or download public data with GEOquery , Geometadb, ArrayExpress)
- ✓ Quality control (ArrayQualityMetrics and affyQCReport)
- ✓ Pre-processing (RMA, gcRMA, MAS5, dChip, and Advanced)
- ✓ Differential expression analysis (limma, SAM, RankProd, and maSigPro for time-course data)
- ✓ Sample size and power analysis (ssize)

What makes Microarray Я US truly unique and very useful among all open access microarray data analysis software are the following:

1. The implementation of several up-to-date Affymetrix custom chip description files (CDF) and probe set re-annotations for both Affymetrix (Dai, Wang et al. 2005; Prieto, Risueno et al. 2008; Risueno, Fontanillo et al. 2010) and Illumina (Du, Kibbe et al. 2007; Barbosa-Morais, Dunning et al. 2010) platforms enables a more accurate and precise microarray data analysis.

2. The versatile results output utility tool enables a speedy and easy generation of input files for over 20 most popular functional analysis software, including Ingenuity Pathways Analysis (Ingenuity Systems, www.ingenuity.com), NextBio (Nextbio, www.nextbio.com), DAVID (Huang da, Sherman et al. 2009; Huang da, Sherman et al. 2009), GSEA-P (Subramanian, Kuehn et al. 2007), GeneTrail (Backes, Keller et al. 2007), WebGestalt (Zhang, Kirov et al. 2005), GeneCodis (Nogales-Cadenas, Carmona-Saez et al. 2009), FatiGO+ (Al-Shahrour, Minguez et al. 2007), ToppCluster (Kaimal, Bardes et al. 2010), TransFind (Kielbasa, Klein et al. 2010), TFactS (Essaghir, Toffalini et al. 2010), GenMAPP2 (Salomonis, Hanspers et al. 2007), Onto-tools Pathway-Express (Draghici, Khatri et al. 2007), FuncAssociate 2 (Berriz, Beaver et al. 2009), GoMiner (Zeeberg, Feng et al. 2003), Gorilla (Eden, Navon et al. 2009), EXALT (Wu, Qiu et al. 2009), The Connectivity Map (Lamb 2007), MAGIA (Sales, Coppe et al. 2010), MMIA (Nam, Li et al. 2009), GeneSet2miRNA (Antonov, Dietmann et al. 2009), and GenePattern (Kuehn, Liberzon et al. 2008), etc. This function facilitates a comprehensive functional analysis of microarray results by drastically cutting down the time and efforts required for converting microarray results files to meet specific format requirements for each of the functional analysis program.

Microarray Я US can be run on all major OS platforms, including Microsoft Windows (XP, Vista, and 7), Apple's Mac OS, and Linux.

CHAPTER 2. DOWNLOAD AND INSTALLATION

2.1 SYSTEM REQUIREMENTS

HARDWARE REQUIREMENTS

- Processor: Minimum--Intel Pentium 4 (or equivalent AMD CPU) 2 GHz 32 bit; Recommended—Intel Core 2 Duo (or equivalent) 2 GHz or higher, 64-bit
- RAM: Minimum—1 GB; Recommended—2 GB or greater (large datasets may require more RAM)
- Hard disk space: 1.5 GB for program installation.

OPERATING SYSTEMS REQUIREMENTS

- Microsoft Windows XP or higher, 32 bit or 64 bit
- Mac OS X
- Linux

2.2 DOWNLOAD AND INSTALLATION INSTRUCTION

2.2.1 INSTRUCTIONS FOR WINDOWS

REQUIRED COMPONENTS:

1. R 2.11.1 (Note: some Bioconductor packages implemented in Microarray Я US are not fully supported in newer versions of R)
2. Rtools 2.11
3. Microarray Я US software

Estimated installation time: 5-20 minutes, depending on the computer configuration.

STEP 1: DOWNLOAD AND INSTALL R

- Download **R 2.11.1 (32 bits)** from http://cran.stat.ucla.edu/bin/windows/base/old/2.11.1/ (Regardless of your Windows OS build, always select the 32-bit R to ensure the correct installation of required packages.)
  Go to control Panel > System to check the Windows build of your PC.

- Double click the .exe file to start the installation wizard.

- On “Select Destination Location” window, **default path** is highly recommended.

- On “Select Components” window, select **Full installation** (Fig. 2-1).
- Complete the installation by accepting all other default settings.

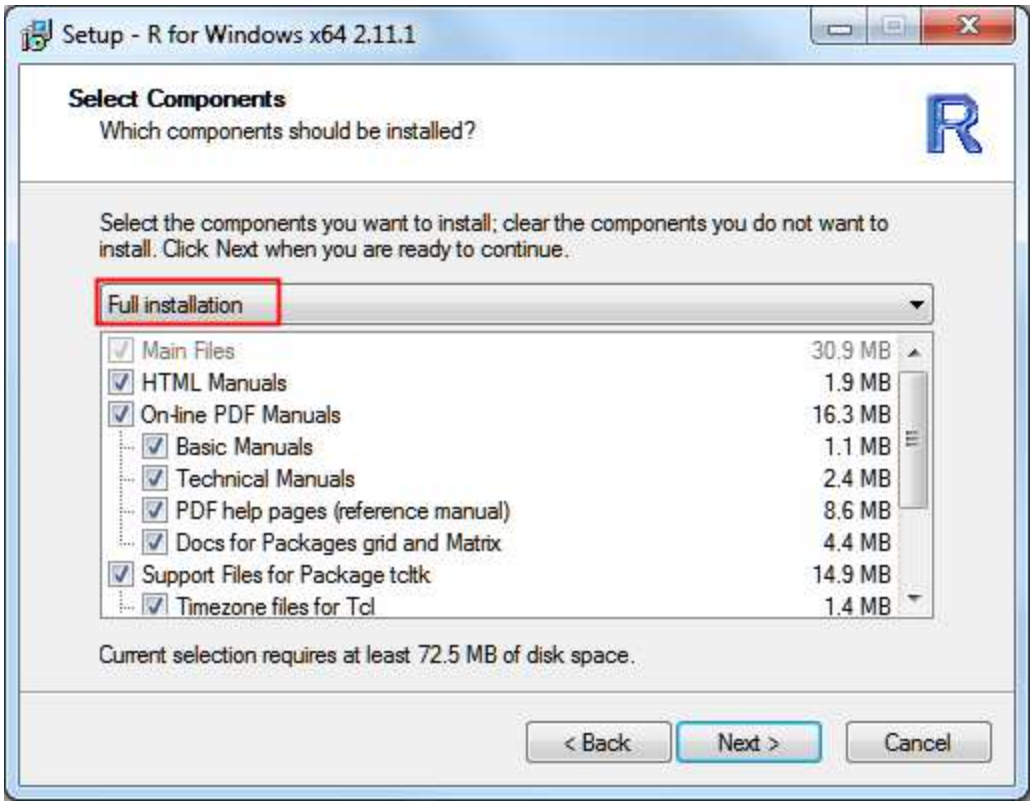


Fig. 2-1 R installation: Select Components window

STEP 2: DOWNLOAD AND INSTALL RTOOLS

- Download the **Rtools211.exe** from http://cran.stat.ucla.edu/bin/windows/Rtools/
- Double click the .exe file to start the installation wizard.
- On “Select Destination Location” window, use the **default path** (C:\Rtools).
- On “Select Components” window, select **Full installation to build R** (Fig. 2-2).
- On “Select R Source Home Directory” window, accept the default path (C:\R).
- On “Select Additional Tasks” window, check all boxes to enable system editing in the next step (Fig. 2-3).
- On “System Path” window, type in the R installation path in the dialogue box (Fig. 2-4).
    - **If R is installed in its default path**, depending on the OS build, type in one of the following:

        C:\Program Files\R\R-2.11.1\bin; For Windows 32 bit
        C:\Program Files (x86)\R\R-2.11.1\bin; For Windows 64 bit

- **If R is NOT installed in its default path**, find out its installation path first.
  - Right click the R desktop icon or start menu shortcut and go to "**Properties**".
  - The R program location is displayed in the "Target" box.
  - The R installation path is the part before bin\.
- **Be sure to append a semicolon to the path name and use forward-slashes (Fig. 2-4).**

- Complete the installation by accepting all other default settings.

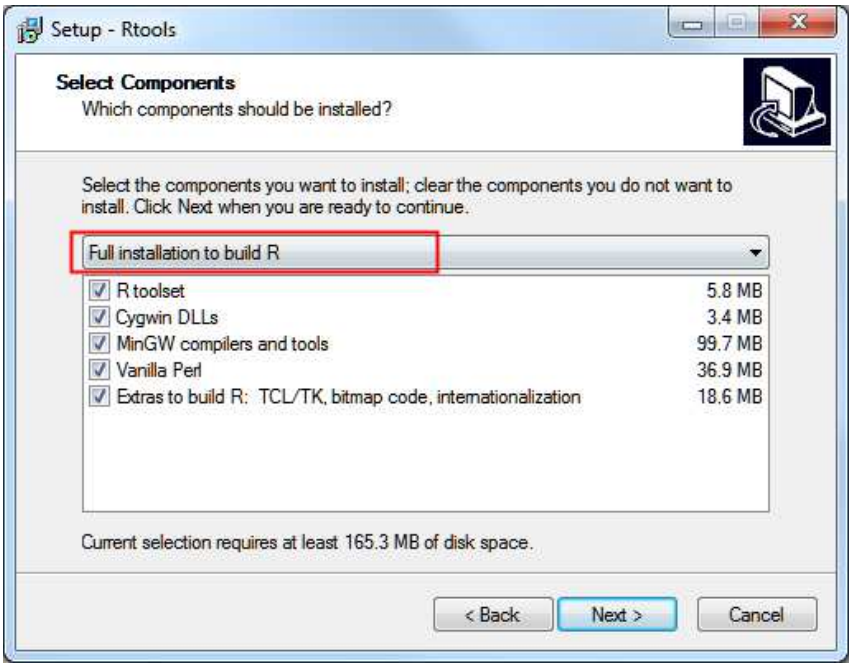


Fig. 2-2 Rtools installation: Select Components

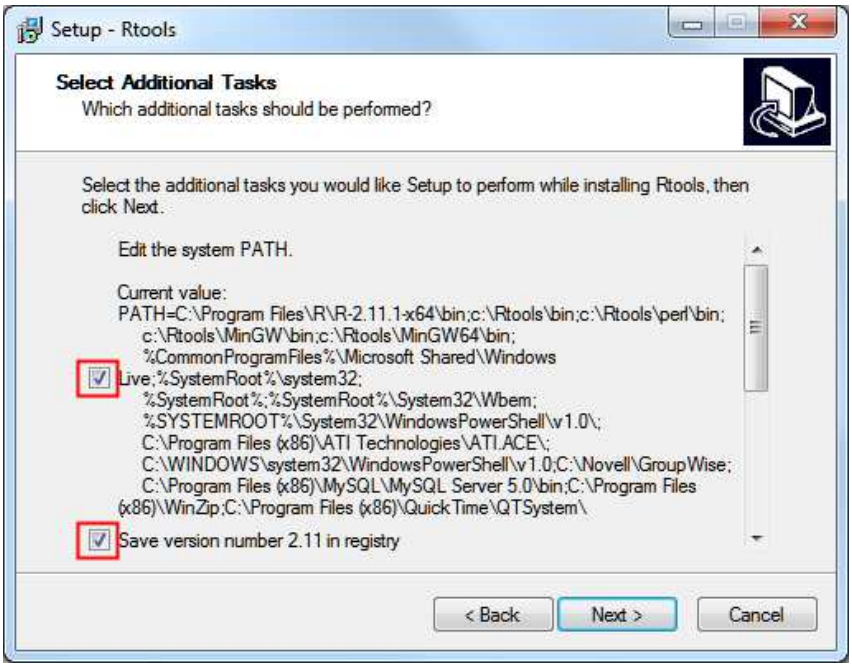


Fig. 2-3 Rtools installation: Select Additional Tasks

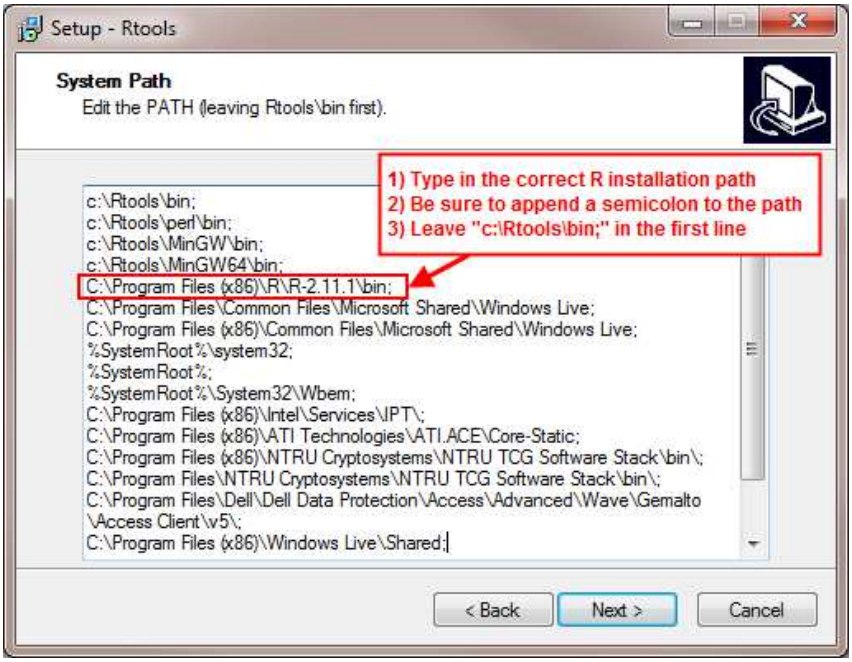


Fig. 2-4 Rtools installation: System Path

STEP3: DOWNLOAD AND INSTALL MICROARRAY Я US

- Microarray Я US download instructions will be sent out upon registration. To register, please follow this link http://norris.usc.libguides.com/aecontent.php?pid=135265&sid=1652613.
- Unzip the downloaded file to **C:**. A folder named **MicroarrayRUS** will be created in the designated path. Within the MicroarrayRUS folder, there should be an R source file **WorkFlow.R** as well as the following four subfolders: **CustomChipBackgroundfile**, **data**, **Install** and **Source**.
- **Make sure the working directory is C:\MicroarrayRUS (i.e. you can find WorkFlow.R file in the following place - C:\MicroarrayRUS\WorkFlow.R).**

2.2.2 INSTRUCTIONS FOR MAC (OS X)

REQUIRED COMPONENTS:

1. X11 (if not pre-installed)
2. R 2.11.1 (Note: some Bioconductor packages implemented in Microarray Я US are not fully supported in newer versions of R)
3. Tck/Tk library for Mac
4. Microarray Я US software

Estimated installation time: 5-90 minutes depending on the computer configuration.

STEP 1: INSTALL X11

- As of OS X 10.5 (Leopard) X11 is installed by default.
  Skip this step if you can find X11 in folder **Applications** > **Utilities** > or **/usr/X11**

- For OS X 10.4 (Tiger), install X11 from the OS X 10.4 installation disk
  - Insert your OS X Tiger Install Disc (#1)
  - Double click on "Optional Installs.mpkg"
  - After selecting the installation drive, expand the "Applications" option and choose "X11" to continue.
  - The X11 application will be installed in /Applications/Utilities/

- For OS X 10.3 (Panther), install X11 from source code
  - Download X11 source code from http://support.apple.com/downloads/X11_for_Mac_OS_X_1_0
  - Double click on the download file, and follow along the installation wizard

- **X11 must be installed BEFORE any other required components.**

> ***I am using OS X 10.5 or higher, but I don't see X11 installed***
> *We suggest you install X11 using the installation disk (similar to OS X 10.4 - Tiger installation above). Otherwise, you need to install the Xcode package, which can be found here: http://developer.apple.com/technologies/tools/xcode.html. Please note that Xcode package is more than **3Gb** in size, and it takes more than **8Gb disk space** and **over 30 minutes** to install.*

STEP 2: DOWNLOAD AND INSTALL R

- Download R 2.11.1 for Mac OS X from http://cran.stat.ucla.edu/bin/macosx/old/R-2.11.1.pkg

- Install R with all default options (double click the downloaded file to install if the installation wizard is not automatically loaded)

STEP 3: INSTALL THE TCL/TK LIBRARY

- Find the latest tcl/tk library for MacOS X from http://cran.stat.ucla.edu/bin/macosx/tools/

- Click the .dmg file (e.g. tcltk-8.5.5-x11.dmg) to download and install (double click the downloaded file if the installation wizard does not automatically load)

- Install the tcl/tk library with all default options

STEP 4: DOWNLOAD AND INSTALL MICROARRAY Я US

- Microarray Я US download instructions will be sent out upon registration. To register, please follow this link http://norris.usc.libguides.com/aecontent.php?pid=135265&sid=1652613.

- Unzip the downloaded file to **/home**. A folder named **MicroarrayRUS** will be created in the designated path. Within the MicroarrayRUS folder, there should be an R source file **WorkFlow.R** as well as the following four subfolders: **CustomChipBackgroundfile**, **data**, **Install** and **Source**.

- **Make sure the working directory is /home/MicroarrayRUS (i.e. you can find WorkFlow.R file in the following place - /home/MiroarrayRUS/WorkFlow.R).**

2.2.3 INSTRUCTIONS FOR LINUX OS

REQUIRED COMPONENTS:

1. R 2.11.1 (Note: some Bioconductor packages implemented in Microarray Я US are not fully supported in newer versions of R)
2. Tcl/tk Table
3. Microarray Я US software

Estimated installation time: 5-20 minutes depending on the computer configuration.

STEP 1: DOWNLOAD AND INSTALL R WITH TCL/TK TABLE

- Install R with tcl/tk packages http://cran.r-project.org/doc/manuals/R-admin.html#Tcl_002fTk

- Special Tktable package (tk package) needed (if tcl/tk<8.5.5). Follow the instructions here: https://stat.ethz.ch/pipermail/r-sig-mac/2006-October/003301.html

- Download the tktable package here: http://sourceforge.net/projects/tktable/files/

STEP 2: DOWNLOAD AND INSTALL MICROARRAY Я US

- Microarray Я US download instructions will be sent out upon registration. To register, please follow this link http://norris.usc.libguides.com/aecontent.php?pid=135265&sid=1652613.

- Unzip the downloaded file to local disk.

- A folder named **MicroarrayRUS** will be created in the designated path. Within the MicroarrayRUS folder, there should be an R source file **WorkFlow.R** as well as the following four subfolders: **CustomChipBackgroundfile**, **data**, **Install** and **Source**.

Special Notes for Linux
Currently, QCReport function is not supported in Linux.

2.3 RUNNING MICROARRAY Я US FOR THE FIRST TIME

2.3.1 ADJUST R MEMORY (FOR WINDOWS ONLY)

Before launching Microarray Я US for the first, adjust R memory to its maximum for best performance. Refer to Table 2-1 for maximum memory allowance in different Windows builds.

- Right click on R desktop icon/start menu shortcut and click on "**Properties**"
- In the "target" box, append " --max-mem-size=?G". Replace "?" with the amount of maximum memory allowed in the operating system, refer to Table 2-1 for details (Fig. 2-5).
- For more information on R memory setting: http://stat.ethz.ch/R-manual/R-devel/library/base/html/Memory-limits.html

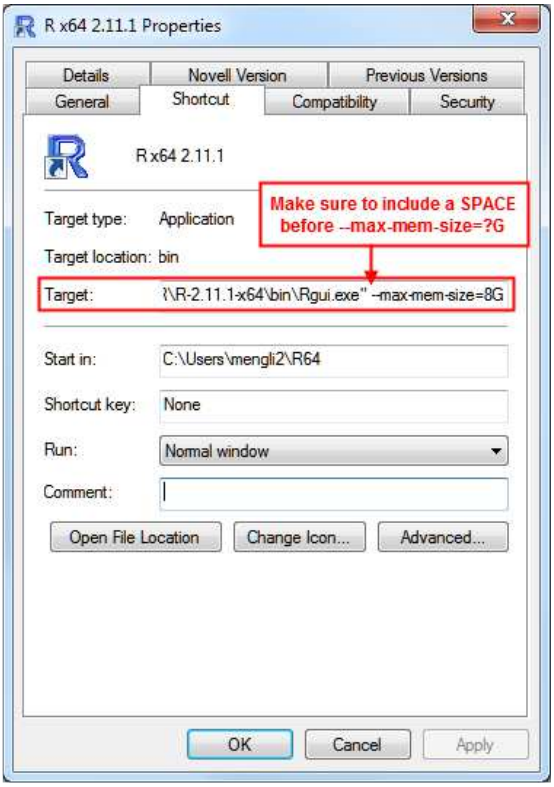


Fig. 2-5 Set R memory in Windows

Maximum Memory Allowance	32-bit Windows	64-bit Windows
32-bit R	The smaller of 2.5 GB and system RAM	The smaller of 3.5 GB and system RAM
64-bit R	Not Applicable	The smaller of 8TB and system RAM

Table 2-1 Maximum memory allowance for R

2.3.2 RUN MICROARRAY Я US FOR THE FIRST TIME

- Make sure the computer is connected to the Internet, preferably via wired Ethernet.
- To open R, right click the R desktop icon/start menu shortcut and select "**Run as administrator**" (Windows) or simply double click the R icon (Mac OS or Linux).
- First select a CRAN mirror by typing the following in the R console:

  chooseCRANmirror()

  Select a CRAN mirror site that is close to your physical location (Fig. 2-6), and click OK.

Fig. 2-6 CRAN mirror window

- Then in the R console, go to File → Open Script (Windows) or Open Document (Mac OS X). Locate the MicroarrayRUS folder and select to open the R source file **WorkFlow.R**.
- WorkFlow.R will now open in R Editor. Edit the Microarray Я US Installation path in the script (Fig. 2-7) and **save the modification** by clicking Ctrl + S on the keyboard or going to File → Save.
    - **For Mac users**, if Microarray Я US was unzipped to the home directory as suggested, the installation path is: ~/MicroarrayRUS

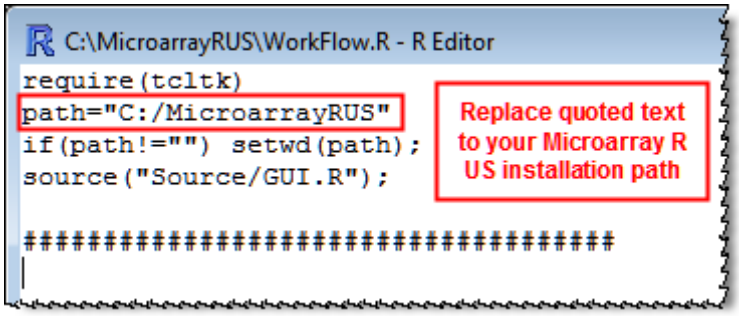


Fig. 2-7 Edit the "WorkFlow.R" script

- To run the script, select all the contents of WorkFlow.R in the R Editor, copy and paste the codes into the R console.
- R will automatically start downloading and installing all the implemented R and Bioconductor packages for Microarray Я US. **This process usually takes 10-30 minutes** (depending on the network speed) and only occurs during the very first run of Microarray Я US.
- Once finished, Microarray Я US console will automatically load (Fig. 2-8).

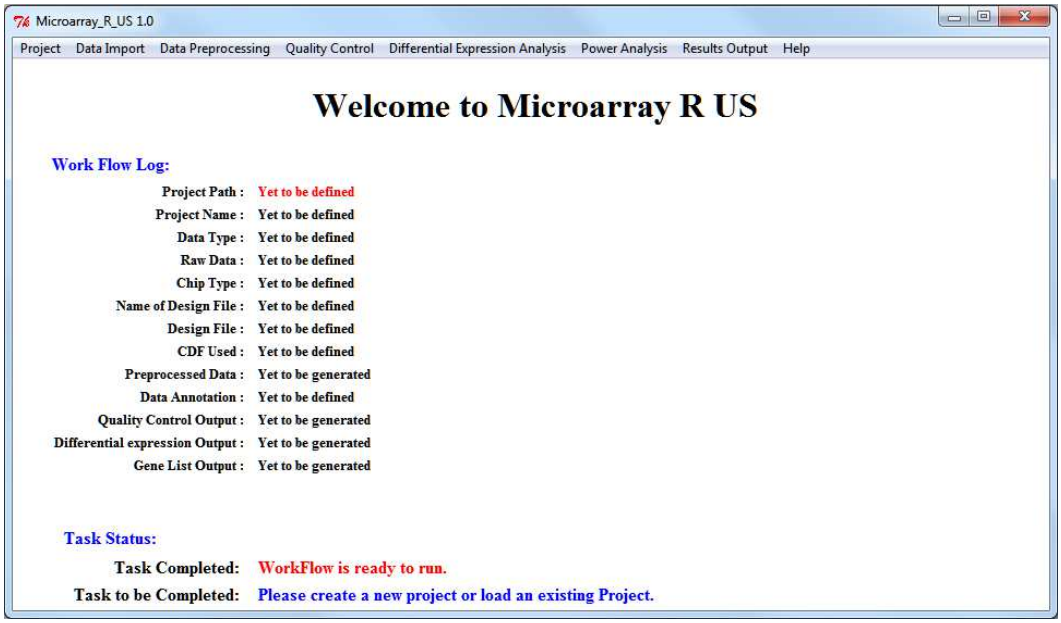


Fig. 2-8 Microarray Я US main window

Future running of Microarray Я US

All R and Bioconductor packages will be installed and ready to use after the first run. To load Microarray Я US in the future, simply open R as administrator and copy-paste the WorkFlow.R script into the R console. There is no need to specify a CRAN mirror or wait for package installation. See Chapter 3.1 for more information.

CHAPTER 3. PROGRAM CONSOLE AND OVERALL WORKFLOW

3.1 PROGRAM CONSOLE

Microarray Я US console is a user-friendly graphic interface. All functionalities can be found on the top navigation bar. The main console display features a **Work Flow Log** that keeps track of analysis stages, and a **Task Status** report that documents the previous task completed and the next task to be completed. To load the Microarray Я US console,

- Right click the R desktop icon/start menu shortcut and select "**Run as administrator**" to open R.
- In the R console, go to File → Open Script. Locate the MicroarrayRUS folder and select to open the R source file **WorkFlow.R** in the R Editor.
- Select all the contents of WorkFlow, copy and paste the codes into the R console.
- Microarray Я US console will automatically start up (Fig. 2-8).

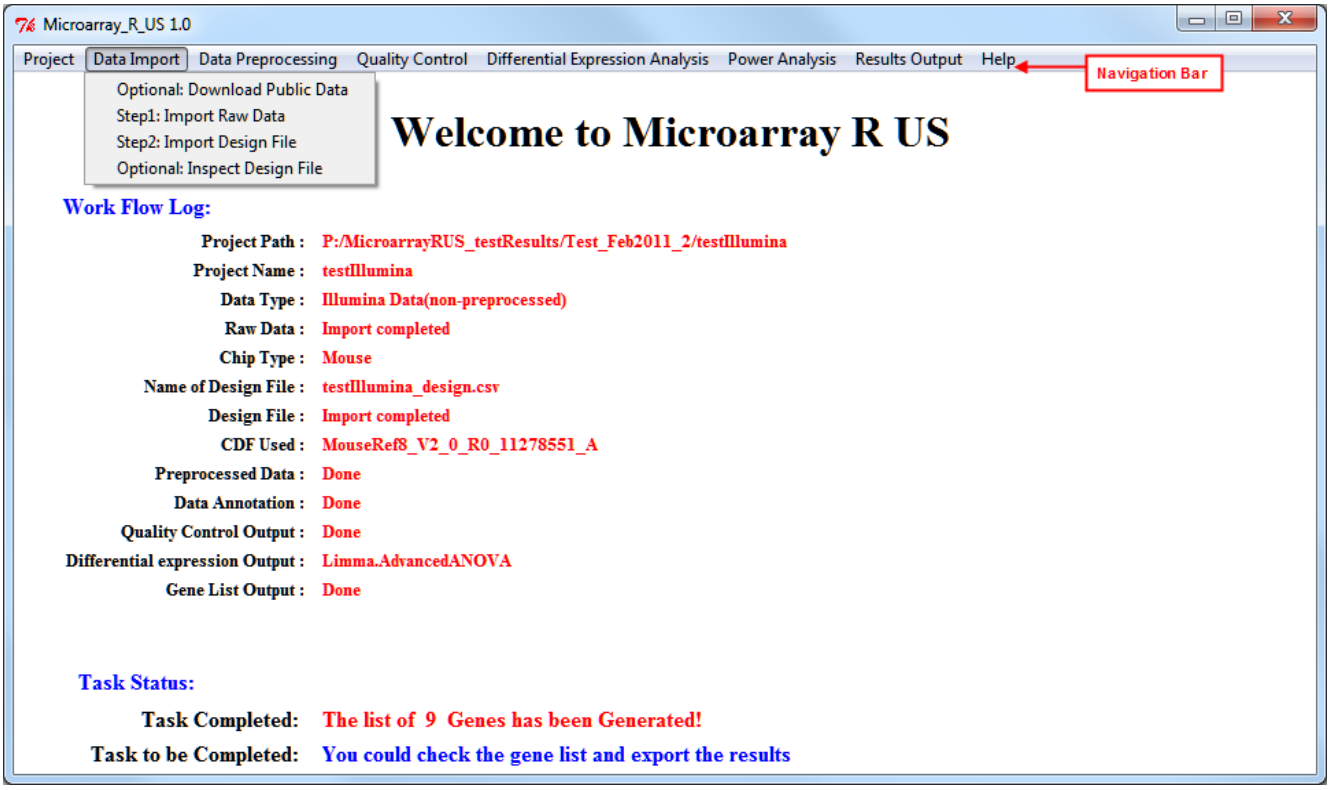


Fig. 3-1 Microarray Я US Console

Error Message "Error in setwd(path) : cannot change working directory"

This error message indicates an incorrect working path setting. Open the WorkFlow.R in R editor (or any text editor, e.g. WordPad). Check and edit the quoted text in line 2 (e.g. path="C:/MicroarrayRUS") to the correct Microarray Я US installation directory (Fig. 2-7). Save the change and reload the WorkFlow.R script in R.

3.2 OVERALL WORKFLOW

Microarray Я US features a linear workflow for analyzing microarray raw data. To move to the next analysis step, the preceding step MUST be completed (except for the Gene List Output Utilities, Venn Diagram and Draw Heatmap of Differentially Expressed Genes, refer to Chapter 10 for details). Fig. 3-2 illustrates the major analysis steps in the Microarray Я US. When using the Microarray Я US, users can simply follow the workflow by going through the Navigation Bar from left to right. Major analysis steps are also clearly marked in the Task Status section. **Task to be Completed** directs users to the next task in the workflow (Fig. 3-1).

Once a project is created in the Microarray Я US, it can be saved at any step and reloaded from that exact step at a later time to resume the workflow.

If desired, users can go back to any previous step in the workflow (e.g. select a different pre-processing method to analyze the data) at any time. In this case, all succeeding steps MUST be re-performed and workflow log will be overwritten. However, any previously outputted results will be maintained in their corresponding output folders under the same project folder.

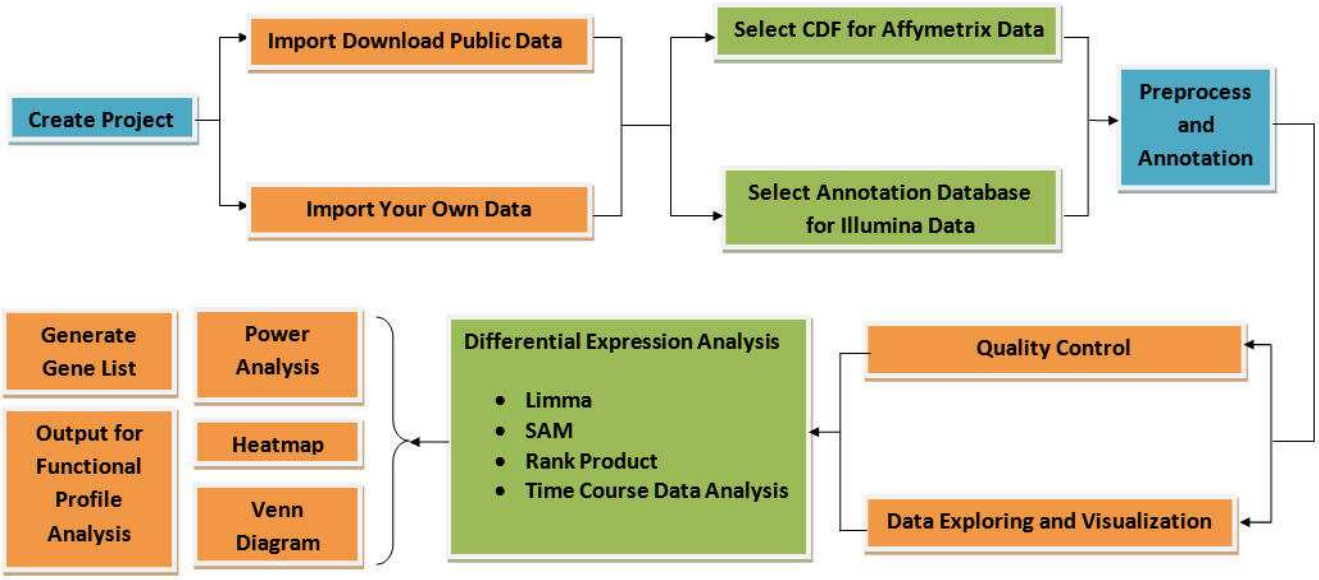


Fig. 3-2 Typical Microarray Я US Workflow

CHAPTER 4. PROJECT MANAGEMENT

Project management creates a new folder to store all results and related R files for the current project. Refer to the **Project** menu in the main Microarray Я US window.

4.1 CREATE A NEW PROJECT

- Click on **Project > Create New Project** to open the "Create New Project" dialogue box
- Specify a directory path to create the new project (Fig. 4-1)

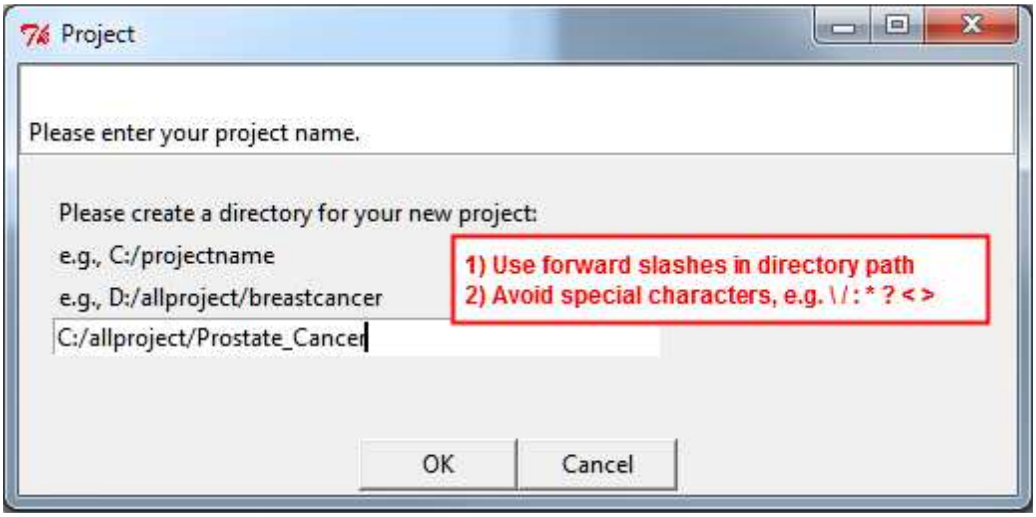


Fig. 4-1 Project window in Microarray Я US

4.2 OPEN EXISTING PROJECT

- Open Existing Project allows you to resume the analysis from the lastly saved progress point
- To open an existing project, click on **Project > Open Existing Project**
- Select a .prj file to load. A .prj files is a log file automatically generated for each project. It stores the project progresses in Microarray Я US.

> ***Project Management Recommendation***
> *Create a main folder (e.g. All Projects) to store all Microarray Я US analyses Projects and subfolders for each experimental data set analysis.*

4.3 SAVE PROJECT

- **Save Project** allows you to document the current analysis progress, including all loaded data loaded and completed analyses.
- To save a project, click on **Project > Save Project**. In Microarray Я US, projects can be saved at any point.
- Files will be saved as .prj files.

Save Project Recommendations

1) ***Save your project frequently.***
2) *Save before closing the Microarray Я US program to retain your most recent analyses.*
3) *Do not edit the saved .prj file; files may not be loaded correctly after editing.*

4.4 OPEN RECENT PROJECT

- **Open Recent Project** allows you to open the five most recent projects processed.
- To open a recent project, click on **Project > Open Recent Project** and select from the list of recently opened projects.

CHAPTER 5. DATA IMPORT

Microarray Я US supports analyses of both user and public data from Affymetrix or Illumina platforms. Data import gets data and design files into Microarray Я US. Refer to the **Data Import** menu in the main Microarray Я US window.

5.1 DOWNLOAD PUBLIC DATA (OPTIONAL)

- **Download Public Data** allows you to download public data from Gene Expression Omnibus (GEO) (Edgar, Domrachev et al. 2002; Barrett, Troup et al. 2011) or ArrayExpress (Parkinson, Sarkans et al. 2011) by dataset IDs (Fig. 5-1).
- Enter the dataset ID and click the "Download" button to start data downloading.
- The downloaded data will be stored in the project\dataset ID folder.

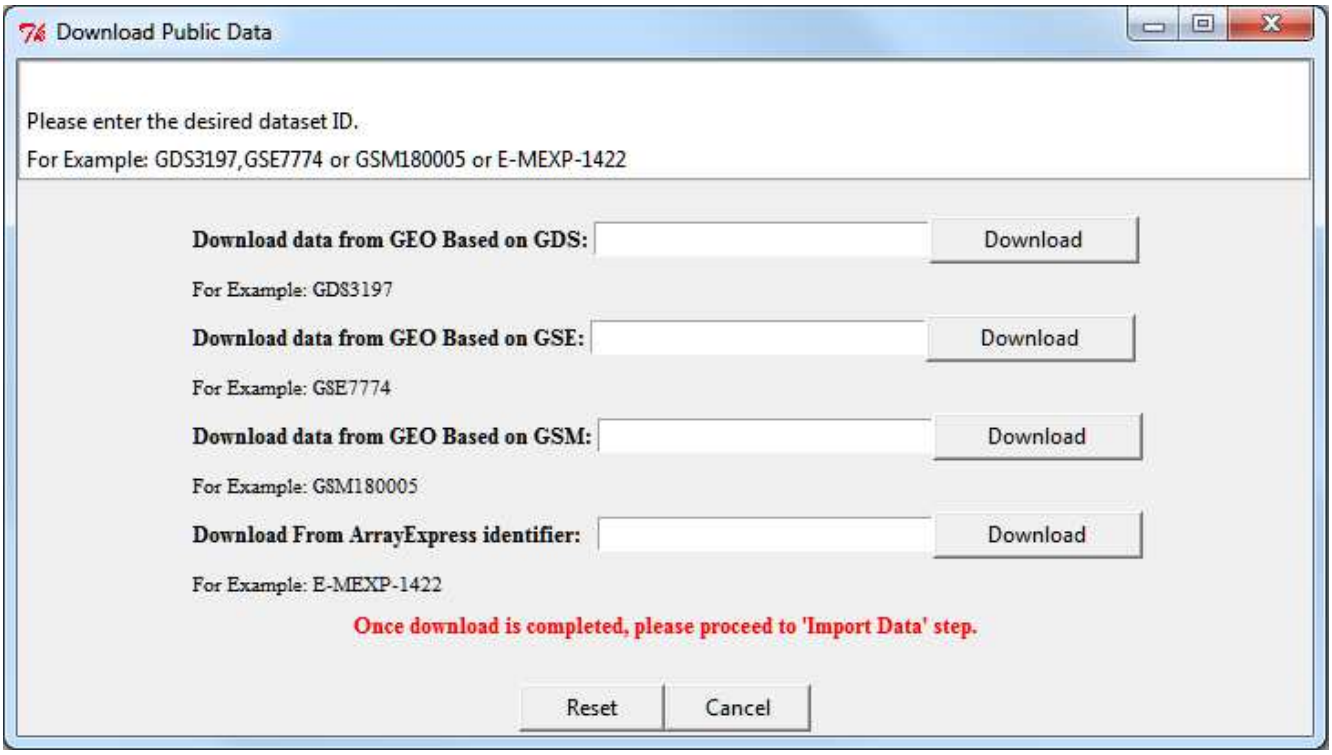


Fig. 5-1 Import Data - Download Public Data

5.2 IMPORT RAW DATA

Data Import allows you to import your own data or previously downloaded public data from local disk. Microarray Я US supports analyses of both Affymetrix GeneChip (CEL files) and Illumina BeadArray data (BeadStudio or GenomeStudio outputs). Click on **Data Import > Step 1: Import Raw Data** to import raw data.

5.2.1 AFFYMETRIX DATA IMPORT

- Select **Affy Data (*.CEL)** in the **Select your data type** dialogue box (Fig. 5-2).
- In the **Import Data** window, select the folder containing all CEL files by **browse**, and specify a design file name or uncheck the option if you have had a design file at hand (Fig. 5-3).

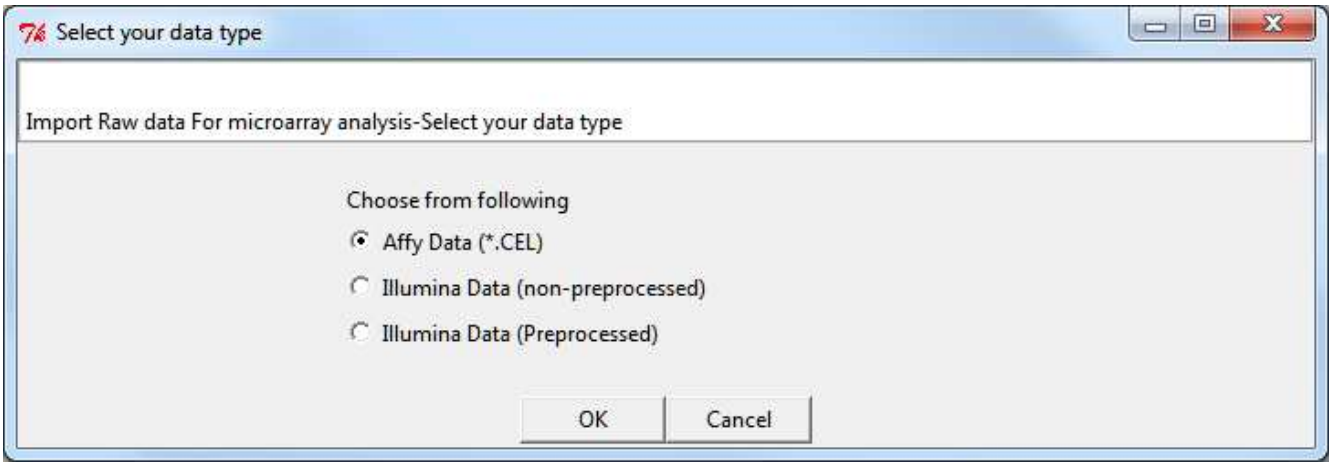


Fig. 5-2 Data Import > Import Raw Data: Select your data type

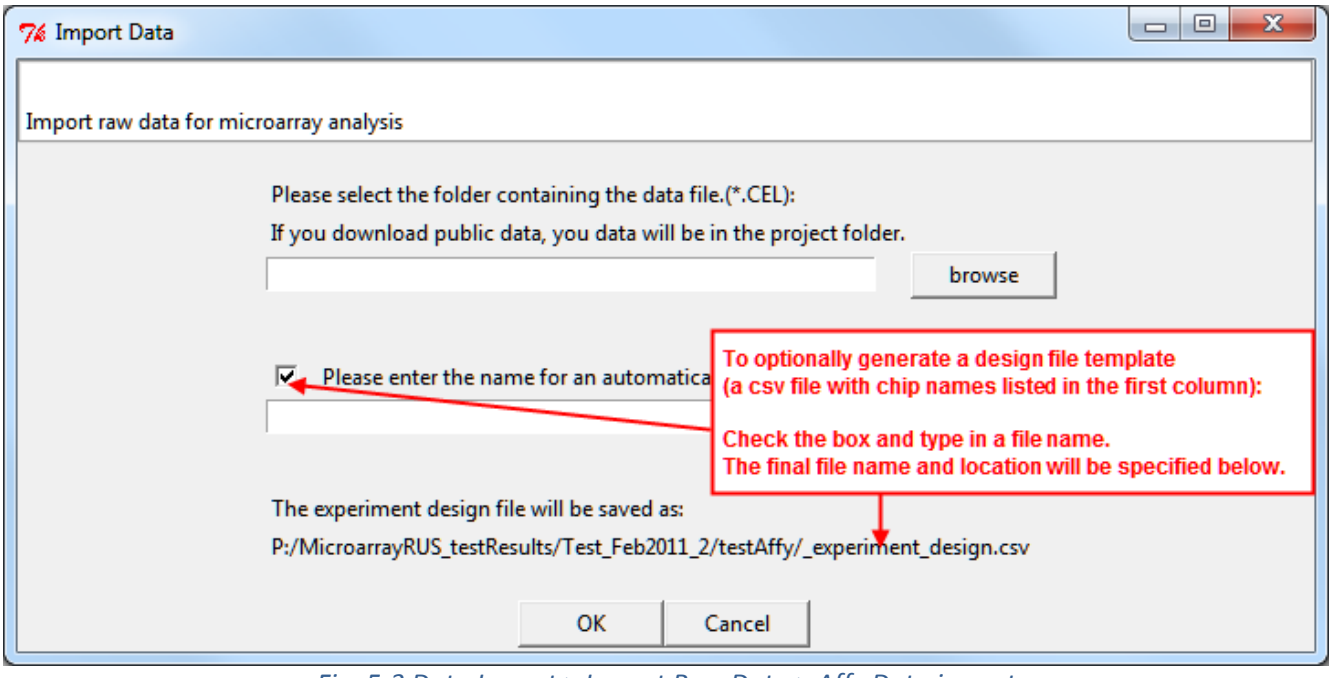


Fig. 5-3 Data Import > Import Raw Data > Affy Data import

Do I need to specify the design file in "Import Data" step?

No, this is an optional function. When specified, data import function generates a design file template with the first column pre-filled with raw data file names. You MUST manually edit the design file to include all sample attributes (experimental factors) in the following columns. If you have your own experimental design file prepared, uncheck the option.

5.2.2 ILLUMINA DATA IMPORT

Microarray Я US supports import of both non-preprocessed (recommended) or preprocessed Illumina data outputted by BeadStudio or GenomeStudio. Please refer to **Appendix 6: Export Illumina gene expression data from BeadStudio for Microarray Я US**.

IMPORT NON-PREPROCESSED ILLUMINA DATA

- If procedures in Appendix 6 were used to export the data and the background control files from BeadStudio, select **Illumina Data (non-preprocessed)** in the **Select your data type** dialogue box (Fig. 5-2).
- In the **Import for Illumina data** window, select the data file and background control file by **browse**, and specify a design file name or uncheck the option if you have your own design file prepared (Fig. 5-4).

IMPORT PREPROCESSED ILLUMINA DATA

- If data has already been preprocessed in BeadStudio, select **Illumina Data (Preprocessed)** in the **Select your data type** dialogue box (Fig. 5-2).
- In the **Import Data** window, select the data file by **browse**, and specify a design file name or uncheck the option if you have your own design file prepared (Fig. 5-5).

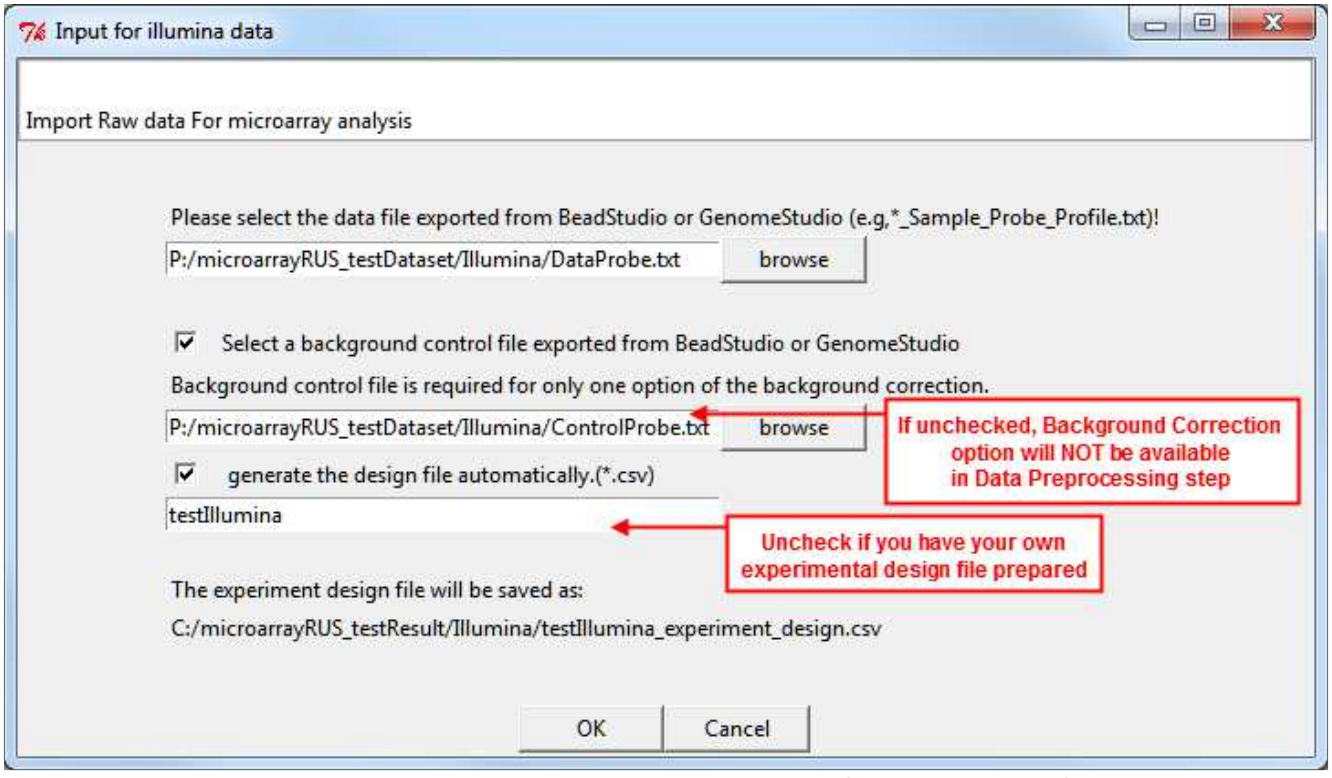


Fig. 5-4 Data Import > Import Raw Data > Illumina data (non-preprocessed) import

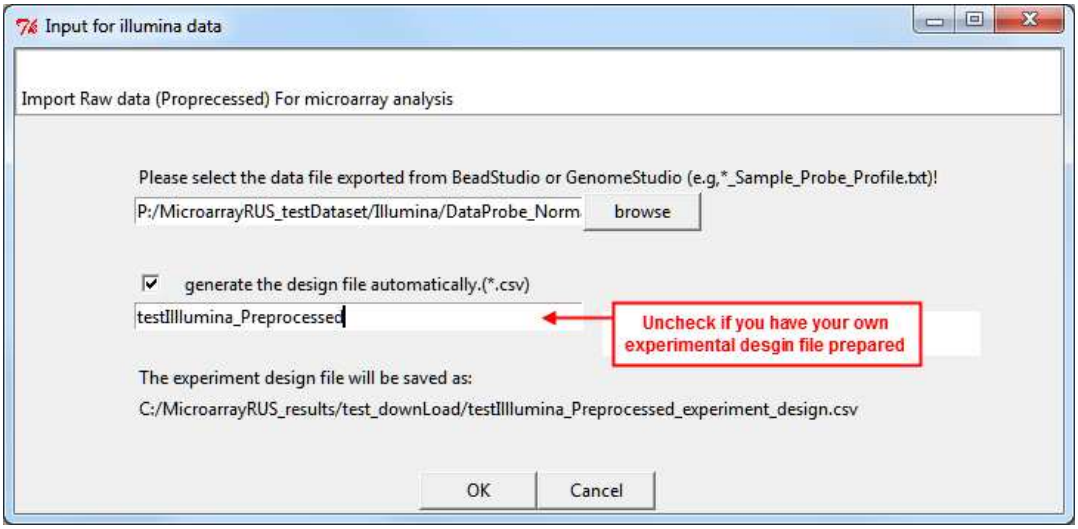


Fig. 5-5 Data Import > Import Raw Data > Illumina data (Preprocessed) import

5.3 IMPORT DESIGN FILE

Once raw data is imported, a **Design File** specifying the experimental set up must be composed.

- If a design file was created in the previous Data Import step, edit the file in Excel to include all related attributes (Refer to **Appendix 7: Notes on Folders and Files**).
- If a custom experimental design file was prepared by user, format the file accordingly and save it as a .CSV file.
- After finish editing, click on **Data Import > Step2: Import Design File** to import the design file.
- Optionally inspect the design file by **Data Import > Optional: Inspect Design File** (Fig. 5-6)

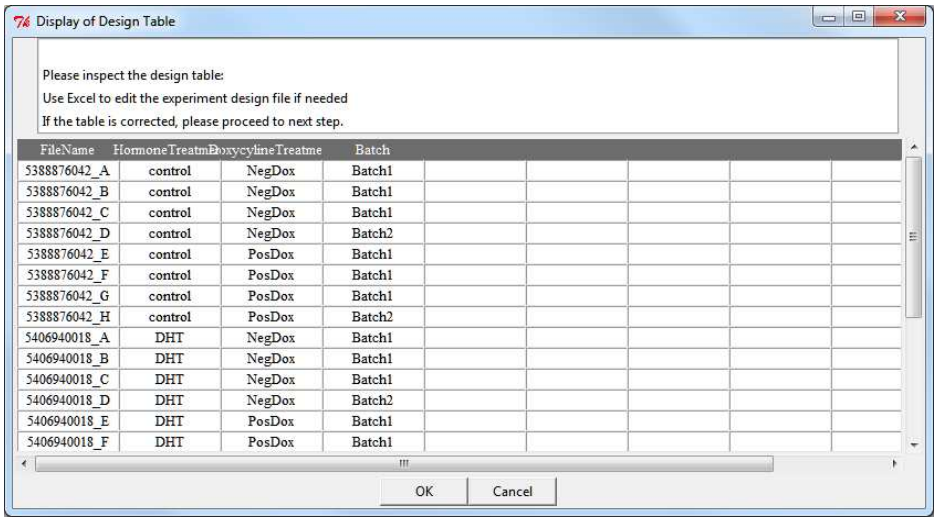


Fig. 5-6 Example: Data Import > Inspect Design File

CHAPTER 6. DATA PREPROCESSING AND ANNOTATION

The aim of the data preprocessing is to remove the technical variances while keeping the biological variations unaffected. Microarray Я US provides several common preprocessing methods, along with advanced user customizable preprocessing methods.

A key feature of Microarray Я US is the implementation of several up-to-date Affymetrix custom chip description files (CDFs) and annotations for both Affymetrix and Illumina platforms, which enables more accurate and precise microarray data analysis.

6.1 SELECT CHIP DESCRIPTION FILE – AFFYMETRIX ARRAY ONLY

Microarray Я US supports Affymetrix's own CDF as well as the custom CDF generated by Dai et al. (version 13) (Dai, Wang et al. 2005) or GATExplorer (Prieto, Risueno et al. 2008). For more details on custom CDFs, refer to **Appendix 4: List of the implemented custom CDF and annotations**. To select a CDF to use, click on **Data Preprocessing > Step1: Select Chip Description File** (Fig. 6-1).

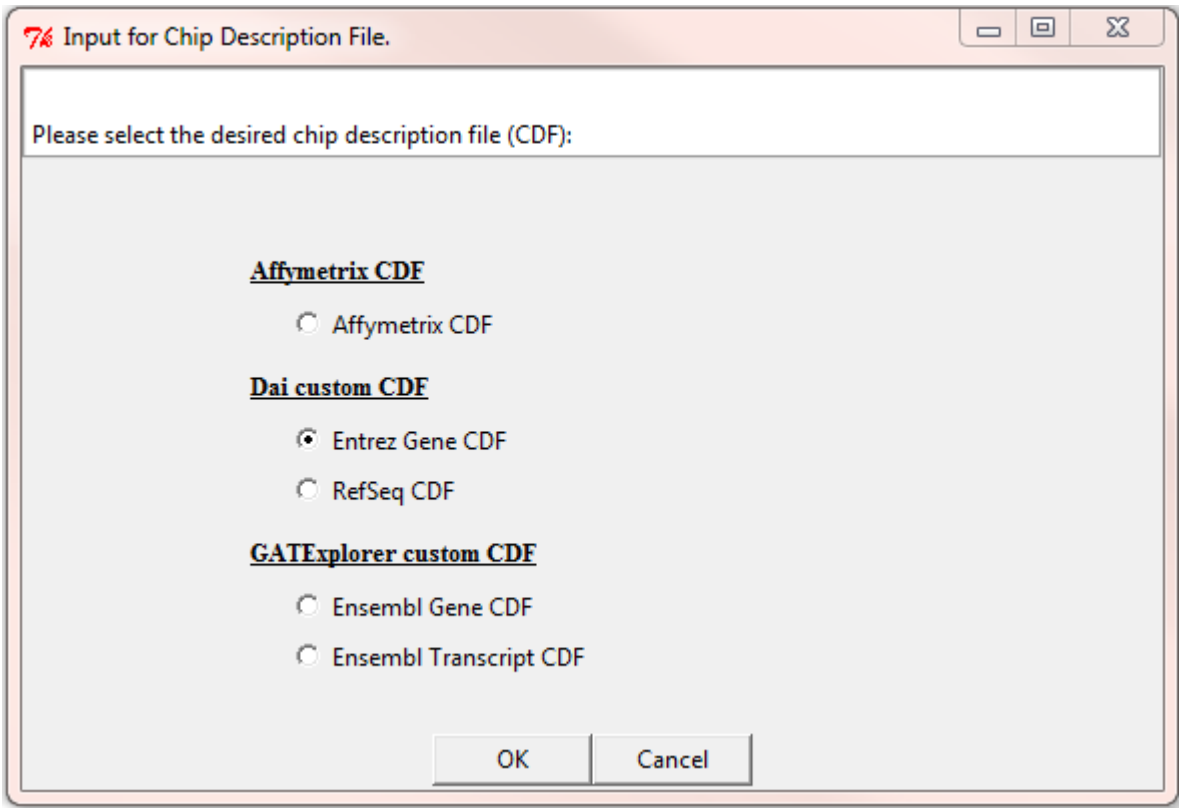


Fig. 6-1 Data Preprocessing > Step1: Select Chip Description File

6.2 PREPROCESS AND ANNOTATE AFFYMETRIX GENECHIP DATA

Once a desired CDF is selected, click on **Data Preprocessing > Step 2: Select Preprocessing Methods** to preprocess Affymetrix data (Fig. 6-2). Microarray Я US provides several common Affymetrix preprocessing methods, including RMA, gcRMA, MAS5.0 and dChip. The **Advanced** option enables customized preprocessing by selecting different algorithms for each preprocessing step, including background correction, normalization, PM correction and summarization (Fig. 6-3). Refer to **Appendix 2: List of the implemented key Bioconductor packages** for detailed descriptions of each algorithm.

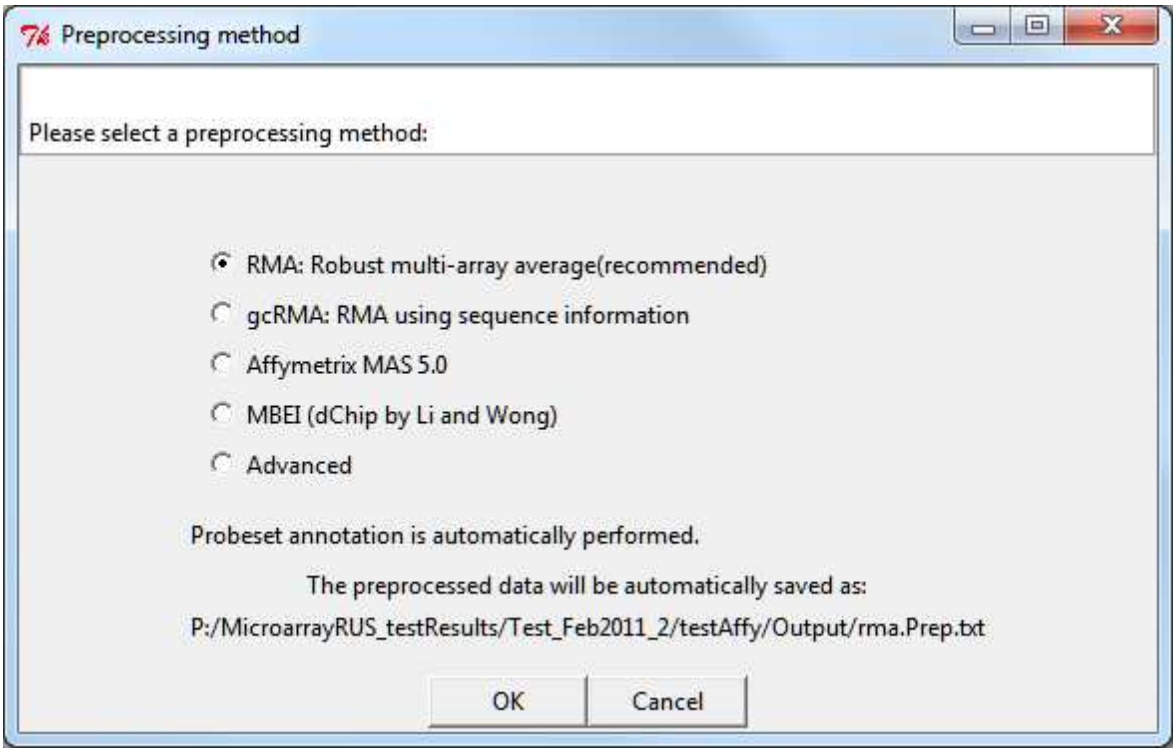


Fig. 6-2 Data Preprocessing > Step 2: Select Preprocessing Methods

SAVE PREPROCESSED AFFYMETRIX DATA

The preprocessed data will be automatically saved as a txt file in the **Output** folder of your project folder. The name of the preprocessed data file will be .Prep.txt prefixed with the selected preprocessing method.

> ***Notes for preprocessed data***
> *Annotations will not be included in the preprocessed .Prep.txt file. Annotations will only be included in output Gene Lists.*

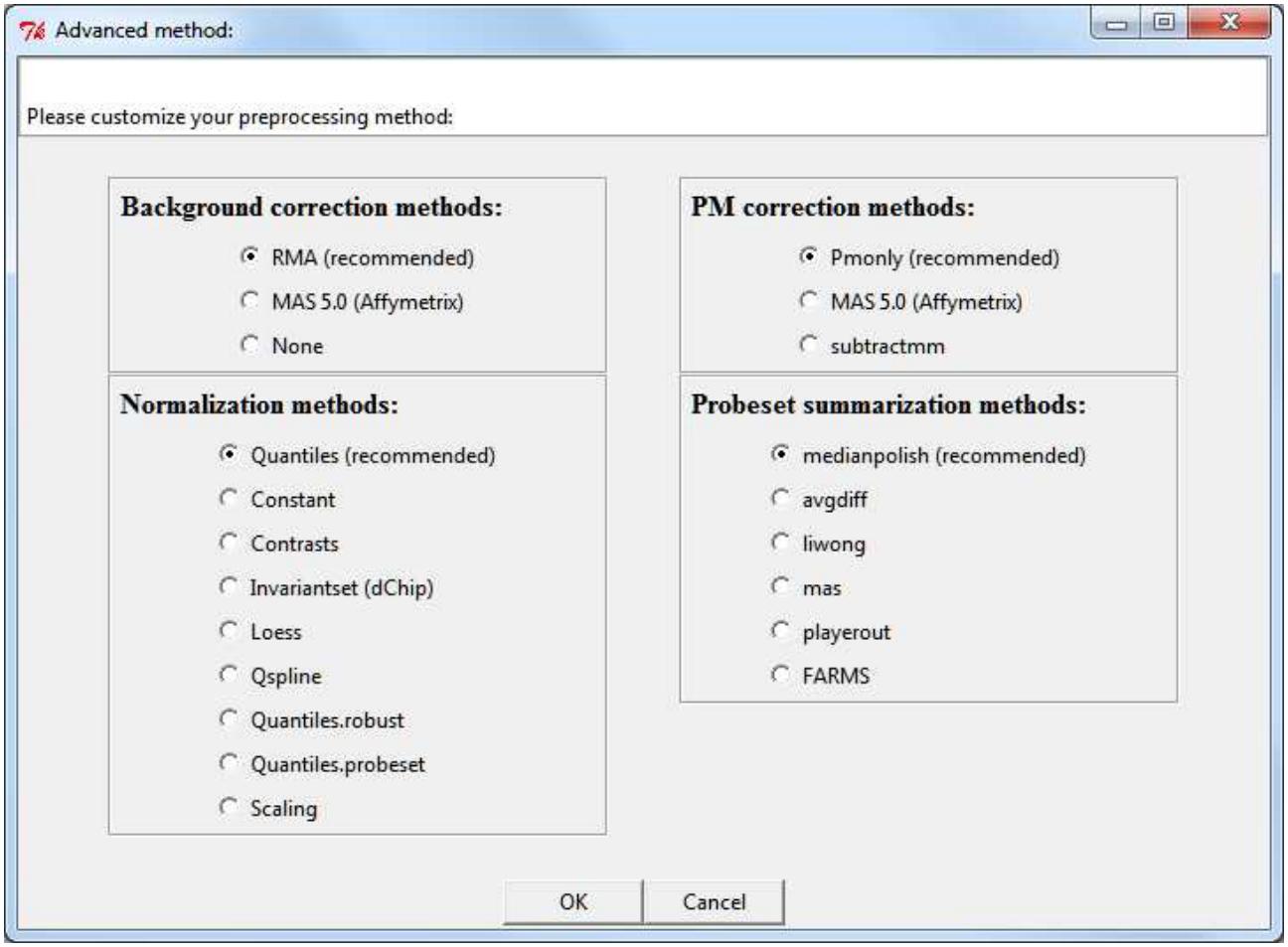


Fig. 6-3 Data Preprocessing > Step 2: Select Preprocessing Methods> Advanced

6.3 PREPROCESS AND ANNOTATE ILLUMINA BEADARRAY DATA

To preprocess and annotate Illumina data, click on **Data Preprocessing > Step 2: Select Preprocessing Methods**. *Note: No custom CDF is available for Illumina data.*

PREPROCESS AND ANNOTATE NON-PREPROCESSED ILLUMINA DATA

For **non-preprocessed Illumina expression data**, Microarray Я US supports a fully customized preprocessing procedure (Fig. 6-4). Users can define the algorithm for each step of the preprocessing, including background correction, variance stabilization and normalization. For details of each algorithm, please refer to **Appendix 2: List of the implemented key Bioconductor packages**.

Three annotations are available for Illumina data: Illumina's own annotation, the re-annotation by Du *et al* (implemented by the Bioconductor lumi package) (Du, Kibbe et al. 2007; Risueno, Fontanillo et al. 2010), and the re-annotation by Barbosa-Morais *et al* (Barbosa-Morais, Dunning et al. 2010). For detailed information regarding the annotations, please refer to **Appendix 4: List of the implemented custom CDF and annotations**.

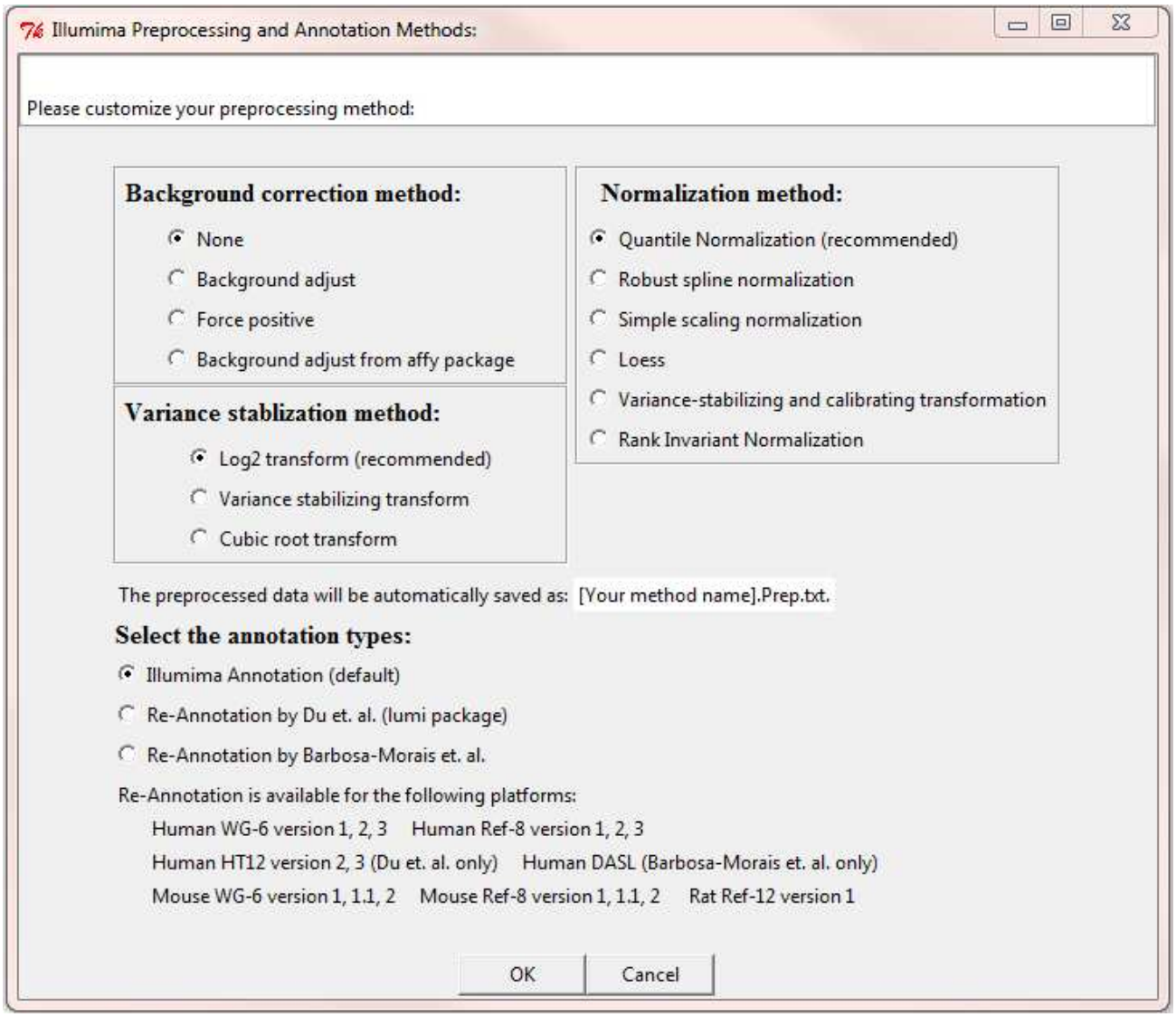


Fig. 6-4 Data Preprocessing > Step 2: Select Preprocessing Methods – Non-preprocessed Illumina data

Chip type selection

- *When using Illumina or Du annotation, a dialogue window will pop up asking if the species is Human. If the model organism is not human, select "No" and then the correct species. Another dialogue window may pop up asking for specific chip type.*
- *When using Barbosa-Morais annotation, a message will pop up if the chip info is not automatically detected, select "Yes" to manually set up the chip info.*

ANNOTATE PREPROCESSED ILLUMINA DATA

For **preprocessed Illumina expression data**, no preprocessing options will be provided. The same annotation options will be provided (Fig. 6-5).

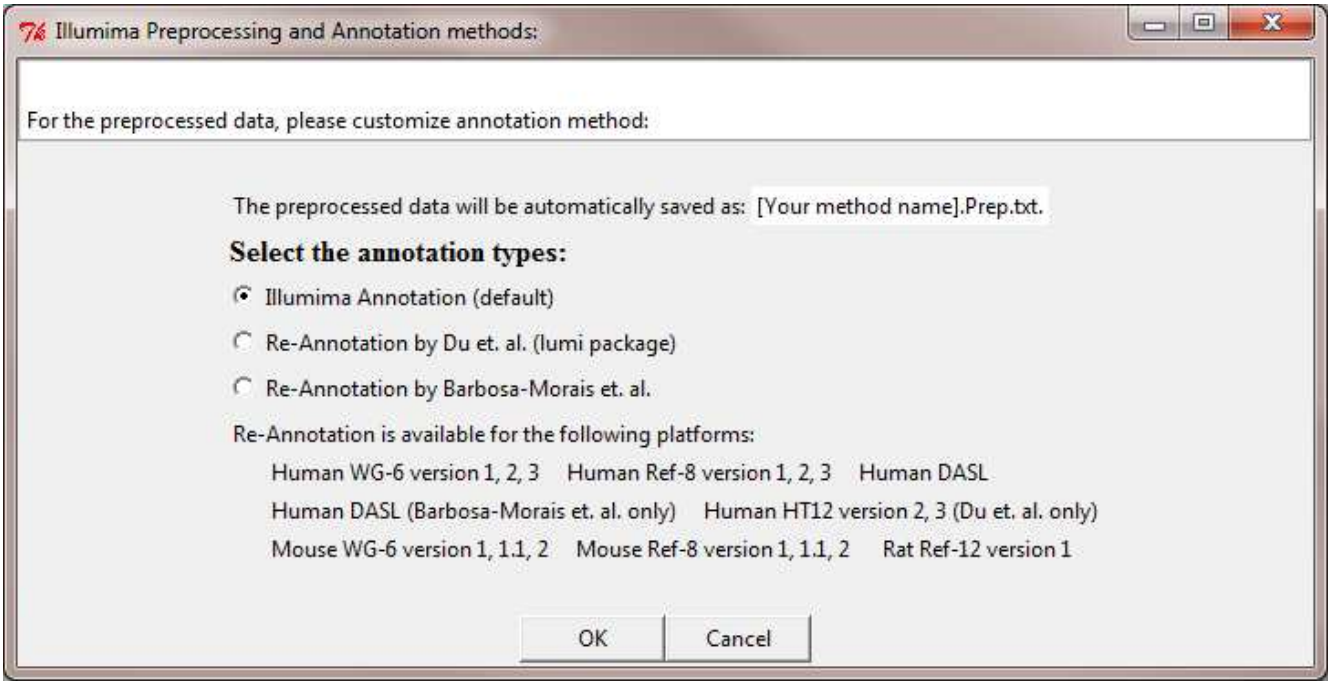


Fig. 6-5 Data Preprocessing > Step 2: Select Preprocessing Methods – Preprocessed Illumina data

SAVE PREPROCESSED ILLUMINA DATA

The preprocessed data will be automatically saved as a txt file in the **Output** folder of your project folder. The name of the preprocessed data file will be .Prep.txt prefixed with the selected preprocessing method.

CHAPTER 7. QUALITY CONTROL AND EXPLORATORY ANALYSIS

Quality control analysis identifies technical artifacts and variance in microarray experiments. Microarray Я US provides convenient and powerful quality control analysis by popular Bioconductor packages for both Affymetrix and Illumina data. Microarray Я US also provides two exploratory techniques, Hierarchical Clustering and Principle Component Analysis (PCA), to quickly examine the global expression patterns across samples.

7.1 GENERATE QUALITY CONTROL REPORT

For Affymetrix data, Microarray Я US implemented two different quality control Bioconductor packages, ArrayQualityMetrics and QCreport. ArrayQualityMetrics algorithm performs an extensive set of quality control testing and generates a summary report (HTML) along with individual PDF plots. QCReport algorithm generates a single PDF report with less quality control testing included. For details, please refer to **Appendix 2: List of the implemented key Bioconductor packages**. Click on **Quality Control > Quality Control** to invoke the quality control dialogue (Fig. 7-1).

For Illumina data, the only supported quality control method is the algorithm implemented in Bioconductor's lumi package. Click on **Quality Control > Quality Control** to start the process.

The results of quality control analysis will be saved in the **QC** folder within your project folder. The results are either in html (ArrayQualityMetrics) or PDF format (QCreport, lumi package).

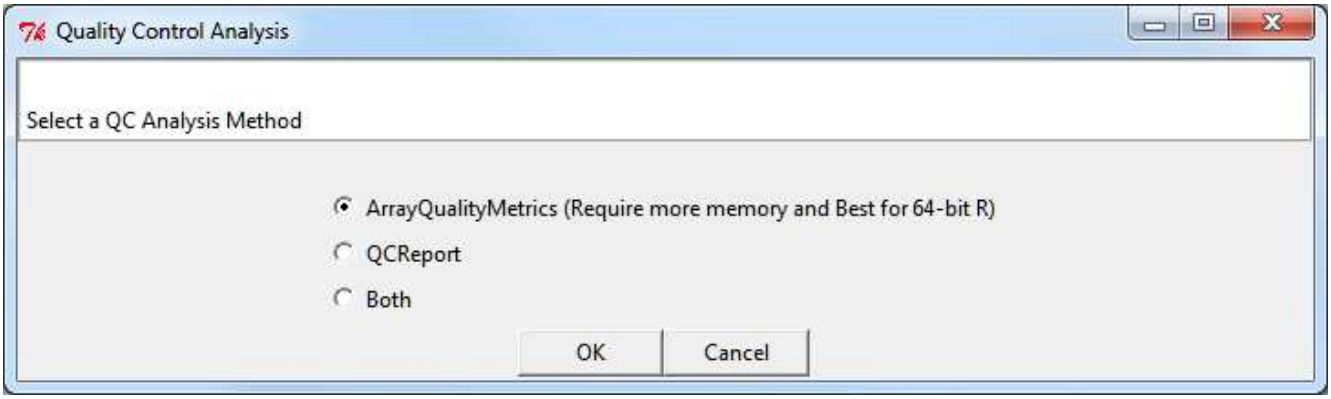


Fig. 7-1 Quality Control > Affymetrix Quality Control Analysis

Notes on using the ArrayQualityMetrics algorithm

ArrayQualityMetrics *is a fairly comprehensive quality control package and requires large memory and long processing time (about 40 minutes to finish a 28-sample Affymetrix Mouse 430 2.0 dataset with R-64 bit on an Intel i7 2.8 GHz-QuadCore with 8GB RAM PC with 64-bit Windows system). We recommend using this option only when you have small dataset or run on a high-performance computer.*

7.2 HIERARCHICAL CLUSTERING ANALYSIS

Hierarchical Clustering analysis of microarray data groups together objects (i.e. genes or samples) with similar expression profiles. Microarray Я US implements various hierarchical clustering and distance measurement algorithms that allow a fully customizable hierarchical clustering analysis (Fig. 7-2). Hierarchical Clustering is performed on EXPERIMENTS ONLY. To invoke the hierarchical clustering dialogue, select **Quality Control > Hierarchical Clustering**.

Clustering results will be saved as a PDF file in the **QC** folder of your project folder.

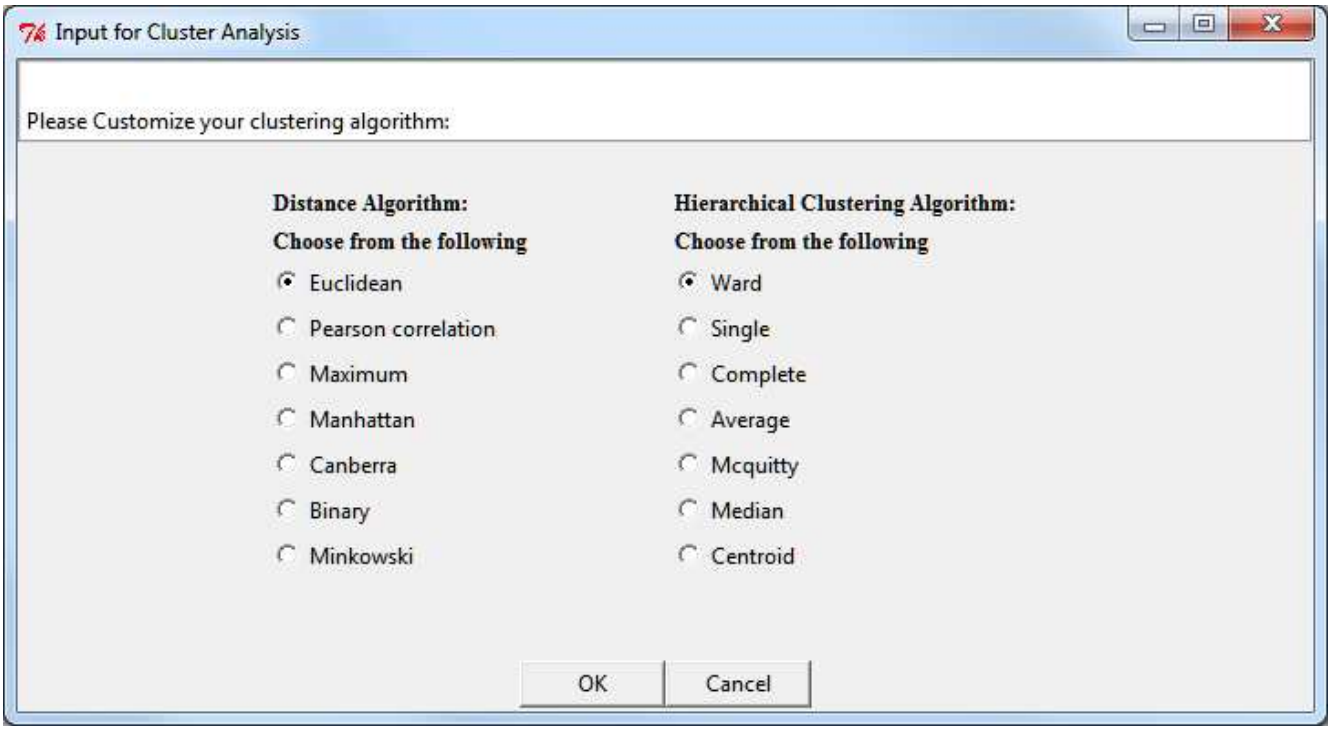


Fig. 7-2 Quality Control > Hierarchical Clustering Analysis

7.3 PRINCIPAL COMPONENT ANALYSIS

Principle Component Analysis (PCA) is a statistical technique for exploring the structure of high dimensional data, such as those generated from microarray experiments. By reducing data dimensionality, PCA allows you to visualize sample relationships in the context of experimental factors, thus infer factors key to the variances in the observations (gene expression). To invoke the PCA function, click on **Quality Control > Principle Component Analysis**. In the PCA configuration dialogue, select an experimental condition to be colored in the PCA plot (Fig. 7-3).

PCA results will be saved as an HTML file in the **QC** folder within your project folder. To view the results, open the report in Internet Explorer and rotate the 3D graph to view from different angles.

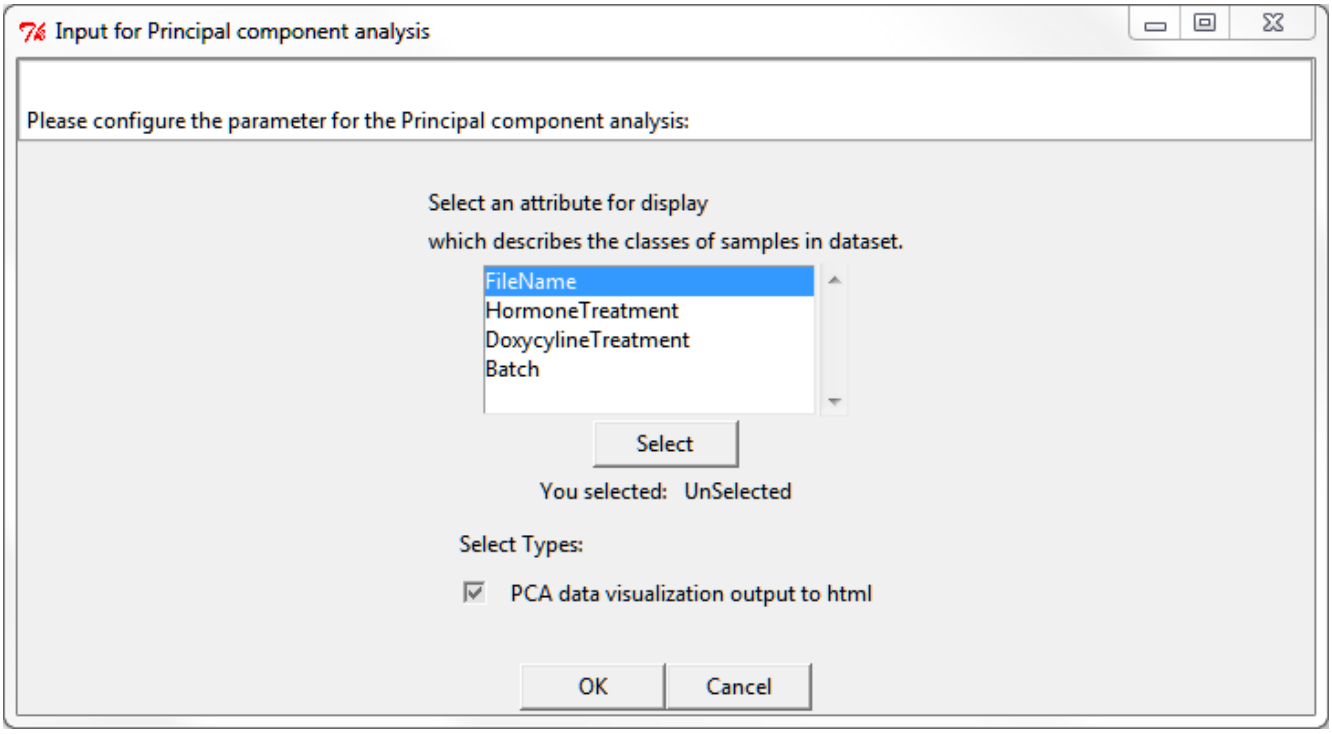


Fig. 7-3 Quality Control > Principle component analysis

CHAPTER 8. DIFFERENTIAL EXPRESSION ANALYSIS

Microarray Я US implements several popular Bioconductor Packages for the statistical analyses of differentially expressed genes from microarray data. They include Linear Model for Microarray Data (LIMMA), Significance Analysis of Microarrays (SAM), Rank Product Test. Refer to **Appendix 3: List of the implemented key methods** for more details

8.1 LINEAR MODEL FOR MICROARRAY DATA (LIMMA)

The linear model method implemented in Microarray Я US can be applied to one factor, two factors, one factor with one random factor, and multiple factors (advanced) experimental designs. To apply LIMMA models, select **Differential Expression Analysis > Limma_Model**.

8.1.1. LIMMA ONE-WAY ANOVA

One-way ANOVA allows users to test one experimental factor for differential expression at a time. The model is most suitable for single factor experiments (e.g. cell type). Select **Differential Expression Analysis > Limma_Model > Limma_1wayANOVA** to invoke the configuration dialogue (Fig. 8-1).

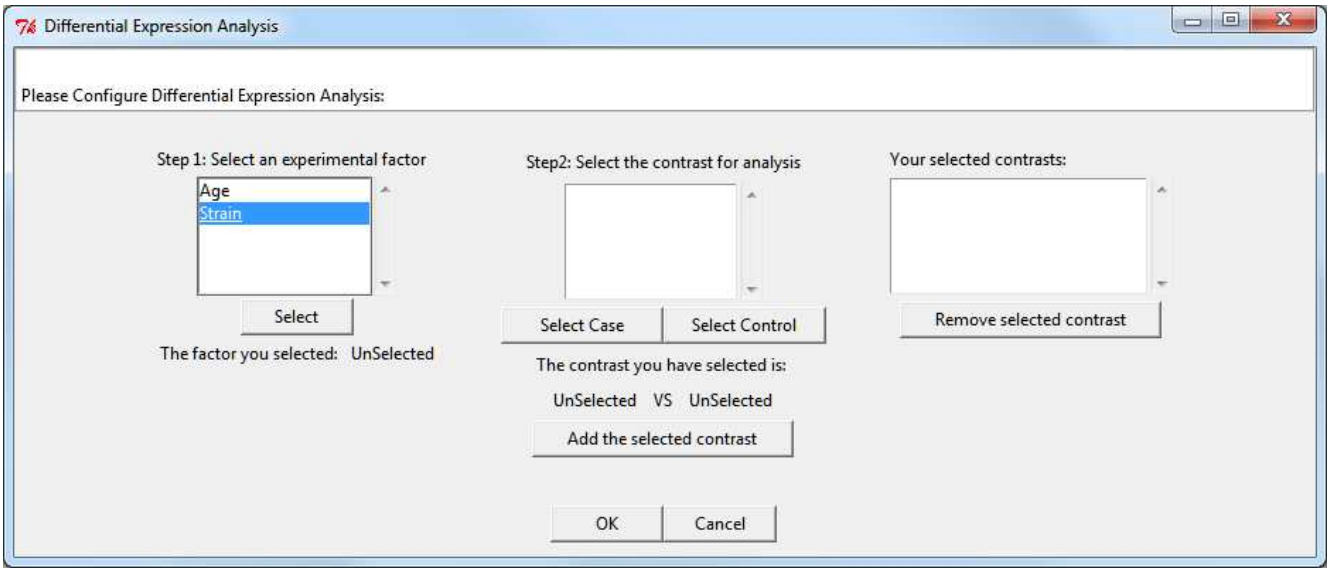


Fig. 8-1 Differential Expression Analysis > Limma_Model > Limma_1wayANOVA

STEP 1: SELECT THE EXPERIMENTAL FACTOR

- All available factors will be listed in **Step 1: Select an experimental factor** box.
- To add an experimental factor, click on the factor to be tested for differential expression and then the **Select** button to choose it.

- Once selected, the text below will change from “The factor you selected: UnSelected” to the chosen experimental factor, for example “The factor you selected: Strain”.
- Available groups of the chosen experimental factor will be automatically listed in **Step 2: Select the contrast for analysis** box (Fig. 8-2).

STEP 2: SELECT THE CONTRAST FOR ANALYSIS

- ANOVA contrast performs a linear comparison of expression values between two specific groups of the factor to generate fold changes. To specify contrast groups, select an available group in the **Step 2: Select the contrast for analysis** box and click on **Select Case** to choose the case group.
- Similarly, select a different group in the **Step 2: Select the contrast for analysis** box and click on **Select Control** to choose the control group.
- Finally, click on **Add the selected contrast** button under the middle box to add the contrast.
- The selected contrast will be listed in the **Your selected contrasts** box (Fig. 8-2).
- Select different groups to add more contrasts.

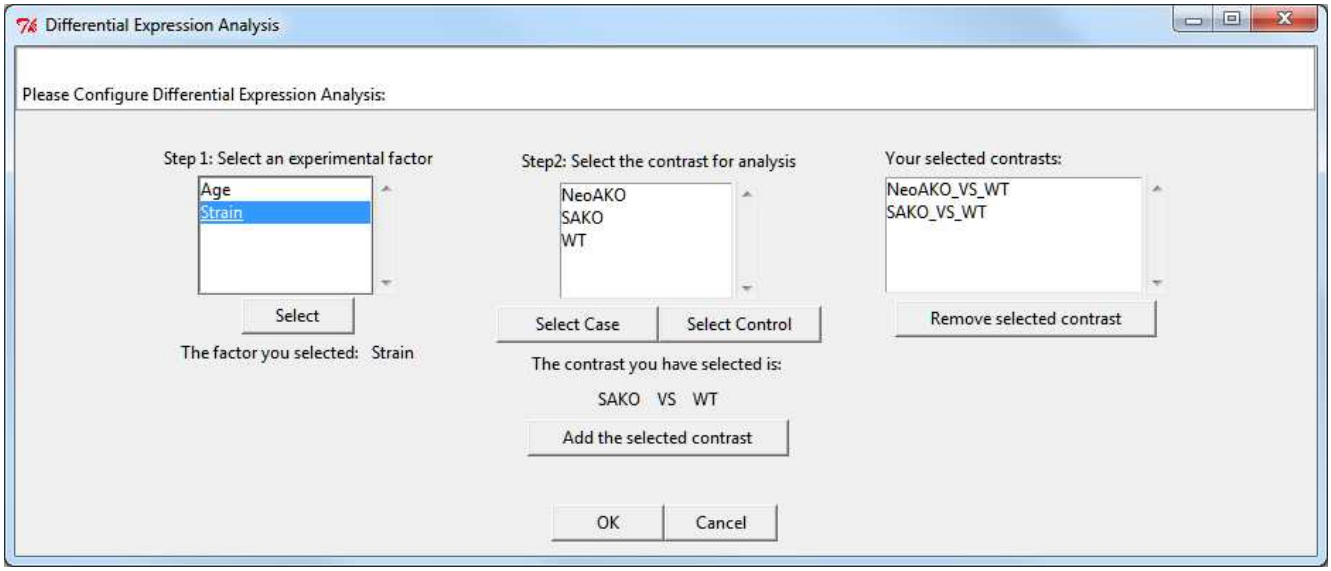


Fig. 8-2 Example: Differential Expression Analysis > Limma_Model > Limma_1wayANOVA

STEP 3: PERFORM DIFFERENTIAL EXPRESSION ANALYSIS

- When finishing adding all the contrasts, click the **OK** button to start the LIMMA one-way ANOVA analysis.

8.1.2. LIMMA TWO-WAY ANOVA

Two-way ANOVA allows users to test two experimental factors and/or the interactions between the two factors for differential expression at a time. The model is most suitable for two factor experiments. Select **Differential Expression Analysis > Limma_Model > Limma_2wayANOVA** to invoke the configuration dialogue (Fig. 8-3).

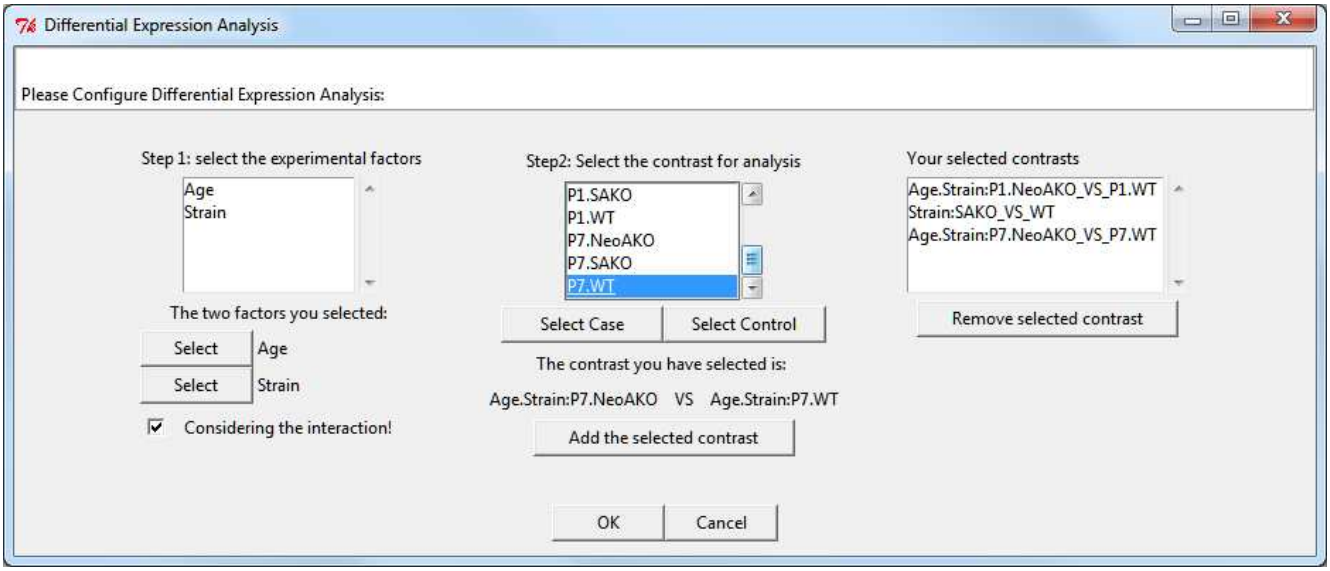


Fig. 8-3 Example: Differential Expression Analysis > Limma_Model > Limma_2wayANOVA

STEP 1: SELECT THE EXPERIMENTAL FACTORS

- All available experimental factors will be listed in **Step 1: Select an experimental factors** box.
- To add experimental factors, one at a time, click on one factor and then the **Select** button to choose it. Two factors are required for two-way ANOVA.
- To add factor interactions, check the **Considering the interaction** box.
- Available groups of the chosen experimental factor and their interactions will be automatically listed in **Step 2: Select the contrast for analysis** box (Fig. 8-3).

STEP 2: SELECT THE CONTRAST FOR ANALYSIS

- Refer to One-way ANOVA Step 2: Select the contrast for analysis

STEP 3: PERFORM DIFFERENTIAL EXPRESSION ANALYSIS

- Refer to One-way ANOVA Step 3: Perform differential expression analysis

8.1.3. LIMMA ONE-WAY RANDOMIZED BLOCK DESIGN

One-way Randomized Block Design allows users to add one experimental factor and one random factor to build the ANOVA model. The model is most suitable for single factor experiments with one random factor (e.g. batch, patient ID). Select **Differential Expression Analysis > Limma_Model > Limma_1wayBlock** to invoke the configuration dialogue (Fig. 8-4).

STEP 1: SELECT THE EXPERIMENTAL FACTOR AND BLOCK (RANDOM FACTOR)

- All available factors will be listed in **Step 1: Select an experimental factor and block** box.
- To add an experimental factor, click on the factor to be tested for differential expression and then the **Select Factor** button to choose it.
- To add a random factor, click on the random factor to be included in the ANOVA model and then the **Select Block** button to choose it.
- Available groups of the chosen experimental factor will be automatically listed in **Step 2: Select the contrast for analysis** box (Fig. 8-4).

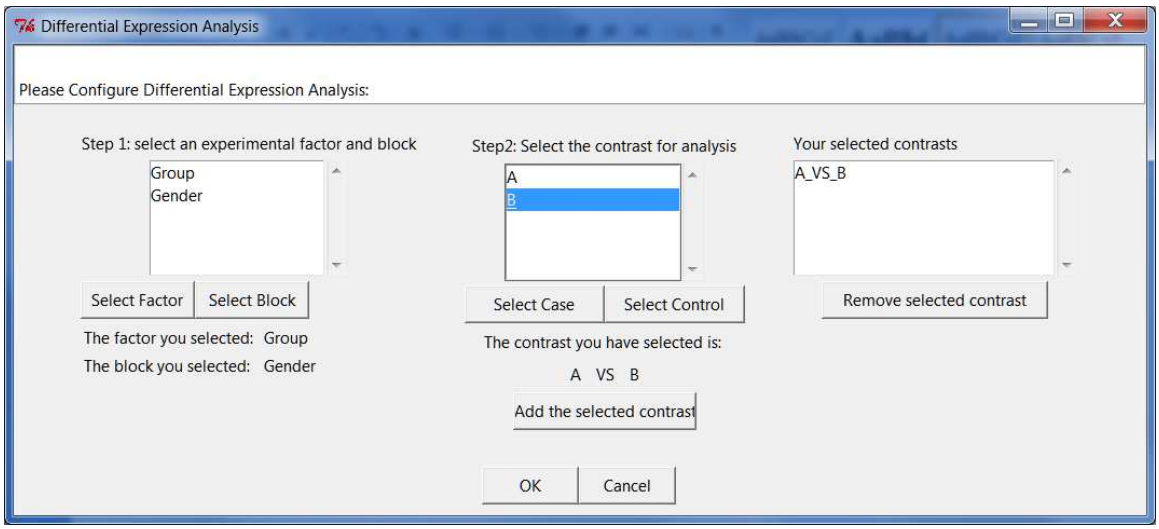


Fig. 8-4 Example: Differential Expression Analysis > Limma_Model > Limma_2wayANOVA

STEP 2: SELECT THE CONTRAST FOR ANALYSIS

- Refer to One-way ANOVA Step 2: Select the contrast for analysis

STEP 3: PERFORM DIFFERENTIAL EXPRESSION ANALYSIS

- Refer to One-way ANOVA Step 3: Perform differential expression analysis

8.1.4. ADVANCED LIMMA MODEL

Advanced LIMMA model allows users to add multiple experimental factors and multiple random factors to build the ANOVA model. The model is most suitable for complicated experimental designs, where multiple experimental factors and their interactions along with random factors are all potentially affecting the gene expression. Select **Differential Expression Analysis > Limma_Model > Limma_Advanced** to invoke the configuration dialogue (Fig. 8-5).

STEP 1: SELECT THE EXPERIMENTAL FACTORS AND BLOCKS (RANDOM FACTOR)

- Available factors will be listed in **Step 1: Select an experimental factor and block** box.
- To add experimental factors, one at a time, click on the factor to be tested for differential expression and then the **Add factor** button to choose it.
- To add random factors, one at a time, click on the random factors to be included in the ANOVA model and then the **Add block** button to choose it.
- The chosen experimental factors will be automatically listed in the **List of the selected experimental factors** box, and random factors in the **List of the selected experimental blocks** box (Fig. 8-5).
- To add interactions between experimental factors, in the **List of the selected experimental factors** box, select one factor then hold the **Ctrl key** to select the second factor and then click the **Add interaction** button to add the interaction between the two factors. To add more interactions between different experimental factors, repeat the step.
- The chosen interactions will be automatically added in the **Step2: Select the contrast for analysis** box.

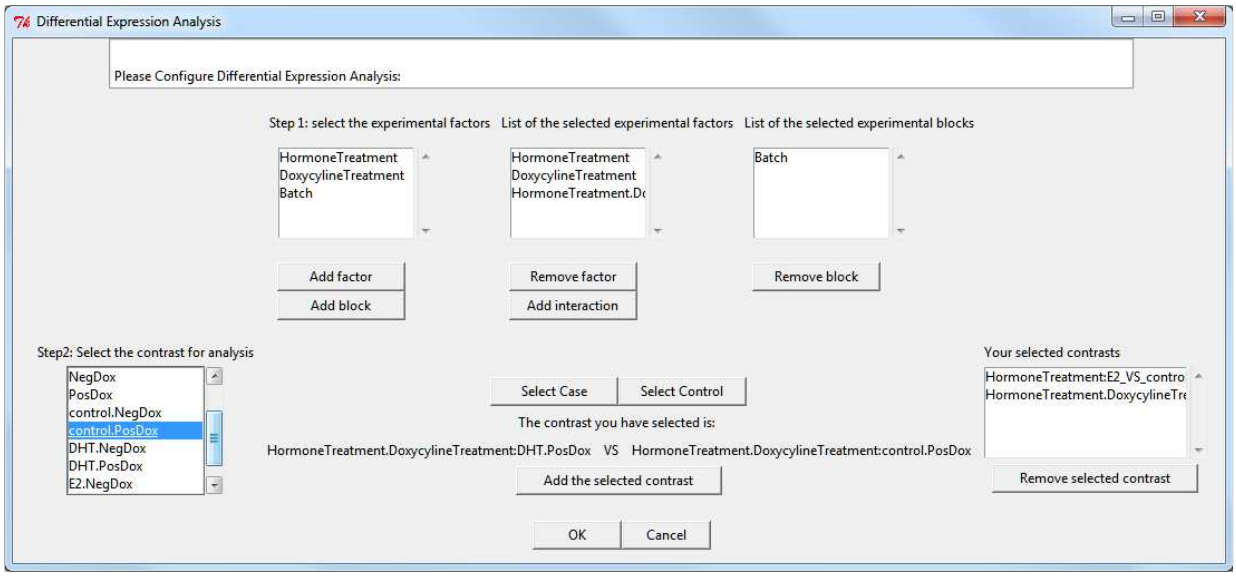


Fig. 8-5 Example: Differential Expression Analysis > Limma_Model > Limma_Advanced

STEP 2: SELECT THE CONTRAST FOR ANALYSIS

- Refer to One-way ANOVA Step 2: Select the contrast for analysis

STEP 3: PERFORM DIFFERENTIAL EXPRESSION ANALYSIS

- Refer to One-way ANOVA Step 3: Perform differential expression analysis

8.2 SIGNIFICANCE ANALYSIS OF MICROARRAYS (SAM)

The SAM method implemented in Microarray Я US can be applied to the single-factor two-group experimental design. For details, please refer to **Appendix 3: List of the implemented key methods**. To perform the SAM analysis, select **Differential Expression Analysis > SAM_Model**.

8.2.1. TWO GROUP UNPAIRED TEST

Two group unpaired test is suitable for two-group experiments with independent samples. Select **Differential Expression Analysis > SAM_Model > SAM_2unpaired** to invoke the configuration dialogue (Fig. 8-6).

STEP 1: SELECT THE EXPERIMENTAL FACTOR

- Available factors will be listed in **Step 1: Select an experimental factor** box.
- Click on the factor to be tested for differential expression and then the **Select** button to choose it.
- Once selected, the text below will change from "The factor you selected: UnSelected" to the chosen experimental factor, for example "The factor you selected: Age".
- Available groups of the chosen experimental factor will be automatically listed in **Step 2: Select the contrast for analysis** box.

STEP 2: SELECT THE CONTRAST FOR ANALYSIS

- SAM performs a linear comparison of expression values between two specific groups for the experimental factor. To specify contrast groups, select an available group in the **Step 2: Select the contrast for analysis** box and click on **Select Case** to choose the case group.
- Similarly, select a different group in the **Step 2: Select the contrast for analysis** box and click on **Select Control** to choose the control group.

STEP 3: CONFIGURE THE PERMUTATION ANALYSIS

- Customize the number of permutations to perform in the SAM analysis (at least 100). This number depends on the size of the user dataset (more permutations for smaller dataset), the expected results accuracy (more permutations for more accurate results), and computer performances (more permutation requires higher performance computers)
- Select a scoring function from **d.stat** or **wilc.stat**. **d.stat** is a modified t-statistics and **wilc.stat** is the Wilcoxon rank test. For details, please refer to **Appendix 3: List of the implemented key methods**.
- Click the OK button to start the SAM analysis.

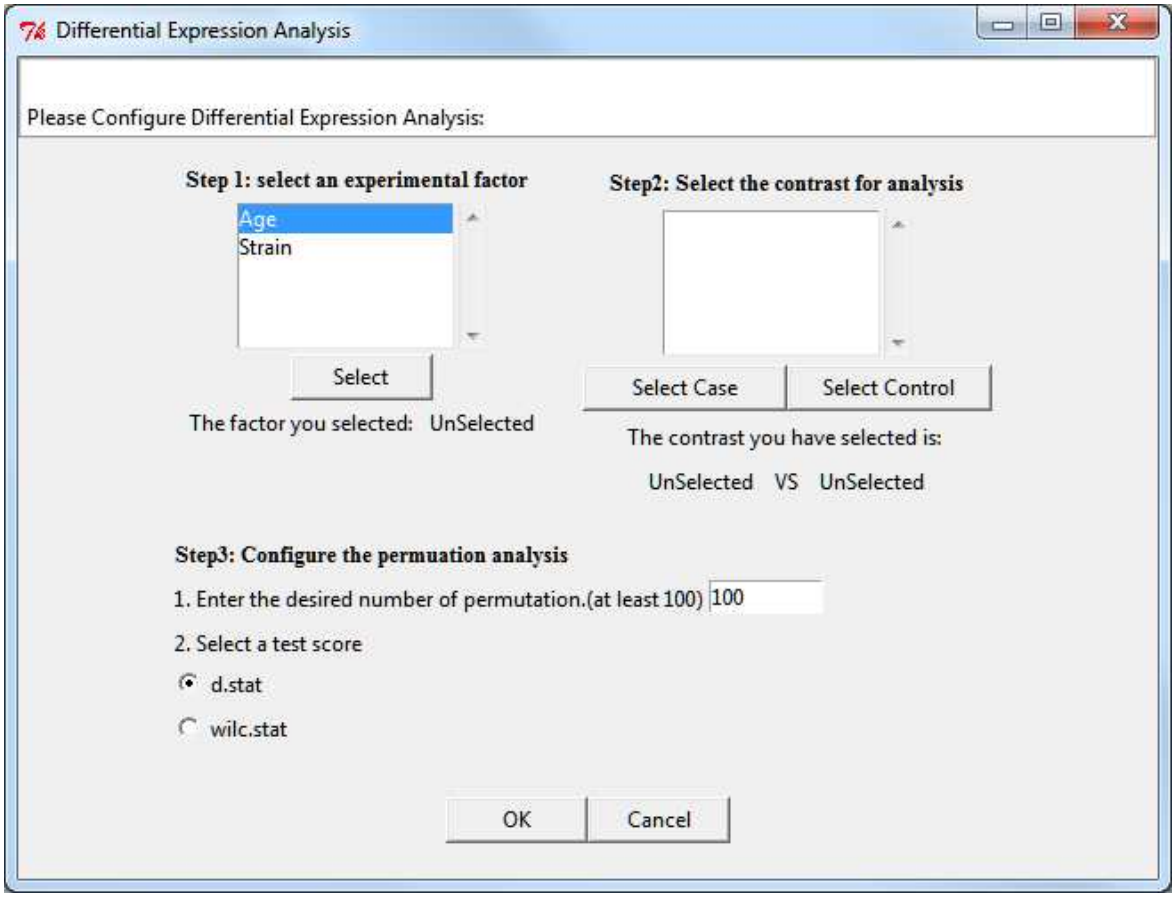


Fig. 8-6 Differential Expression Analysis > SAM_Model > SAM_unpaired

8.2.2. TWO GROUP PAIRED TEST

Two group paired test is suitable for two-group experiments with dependent samples. Select **Differential Expression Analysis > SAM_Model > SAM_2paired** to invoke the configuration dialogue (Fig. 8-7).

STEP 1: SELECT THE EXPERIMENTAL FACTOR

- Available factors will be listed in **Step 1: Select an experimental factor** box.
- Click on the factor to be tested for differential expression and then the **Factor** button to choose it.
- Click on the factor indicating the paired samples (e.g. patient ID) and then the **Paired Vector** button to choose it.
- Available groups of the chosen experimental factor will be automatically listed in **Step 2: Select the contrast for analysis** box.

STEP 2: SELECT THE CONTRAST FOR ANALYSIS

- Refer to SAM two group unpaired test Step 2: Select the contrast for analysis.

STEP 3: CONFIGURE THE PERMUTATION ANALYSIS

- Refer to SAM two group unpaired test Step 3: configure the permutation analysis.

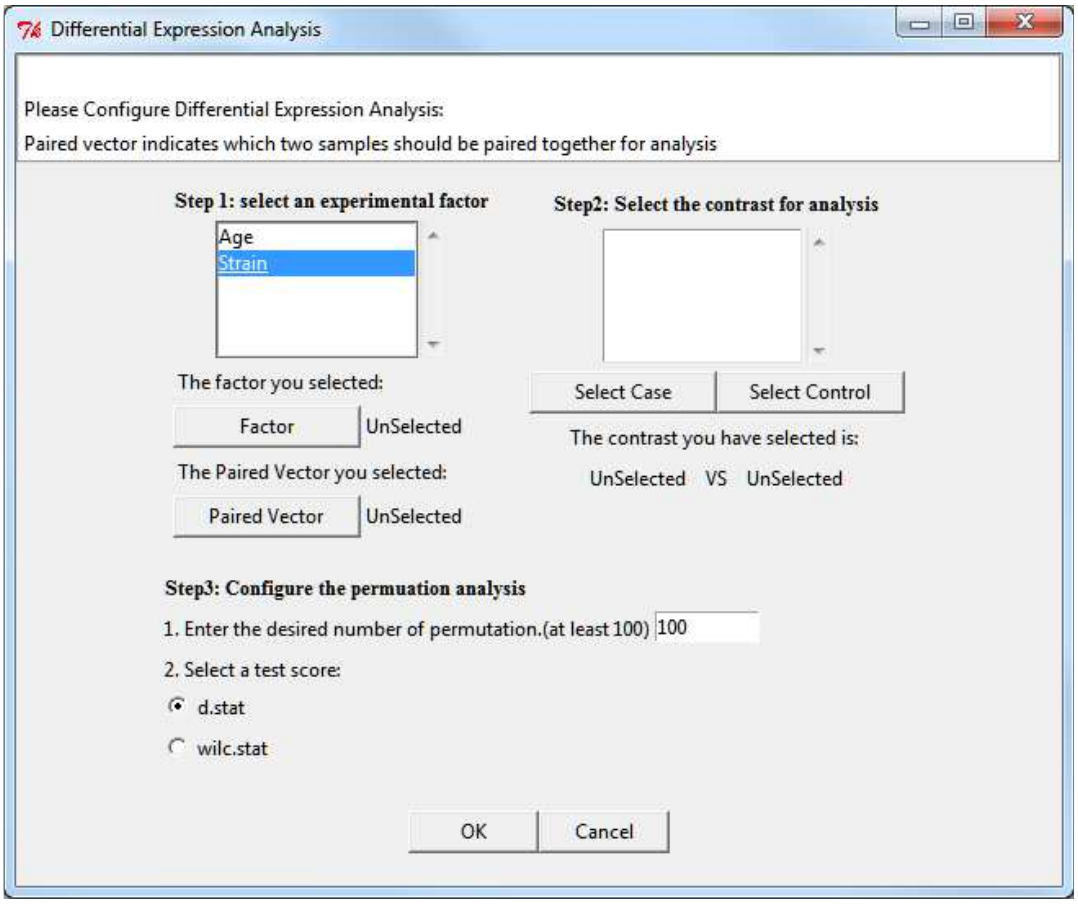


Fig. 8-7 Differential Expression Analysis > SAM_Model > SAM_paired

8.3 RANK PRODUCT TEST

Microarray Я US implements Bioconductor's RankProd package for two-group experiments and also meta-analysis of data from different sources. The rank product test is a non-parametric statistical method based on the rankings of the fold changes of genes. For details about rank product test, please refer to **Appendix 3: List of the implemented key methods**.

8.3.1. RANK PRODUCT TEST (ONE ORIGIN)

One origin rank product test is suitable for two-group experiments. Select **Differential Expression Analysis > RankProduct_Model > RankProd _OneOrigin** to invoke the configuration dialogue (Fig. 8-8).

- Select the experimental factor from the available factor listed in **Step 1: Select an experimental factor** box, and click on the **Select** button to choose it.
- Select the contrasts from the available groups listed in the **Step 2: Select the contrast for analysis** box, and click on **Select Case** to choose the case group and **Select Control** the control group.
- Customize the number of permutations for the one origin rank product test (at least 100). This number depends on the size of the user dataset (more permutations for smaller dataset), the expected results accuracy (more permutations for more accurate results), and computer performances (more permutation requires higher performance computers)
- Click "**OK**" to start the analysis.

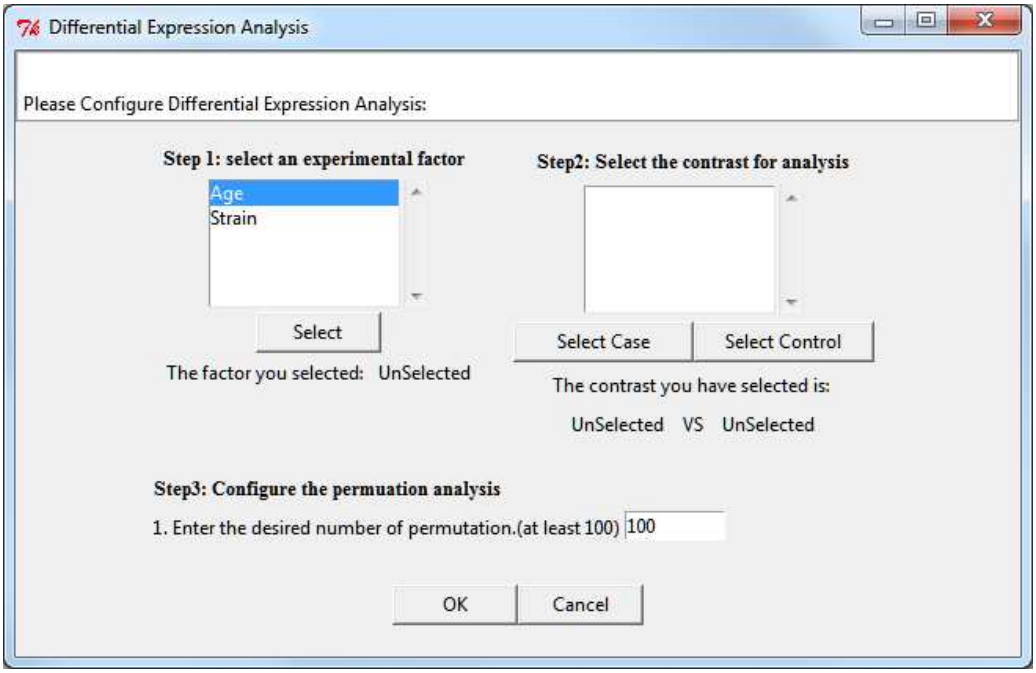


Fig. 8-8 Differential Expression Analysis > RankProduct_Model > RankProd_OneOrigin

8.3.2. RANK PRODUCT TEST (MULTI ORIGIN)

Multi-origin rank product test is the meta-analysis for microarray datasets generated from different experiments or labs. Select **Differential Expression Analysis > RankProduct_Model > RankProd _MultiOrigin** to invoke the configuration dialogue (Fig. 8-9).

- Select the experimental factor from the available factor listed in **Step 1: Select an experimental factor** box, and click on the **Factor** button to choose it.
- Select the experimental factor that specifies sample origins (e.g. experiment accession number) from the available factor listed in **Step 1: Select an experimental factor** box, and click on the **Origin ID** button to choose it.
- Select the contrasts from the available groups listed in the **Step 2: Select the contrast for analysis** box, and click on **Select Case** to choose the case group and **Select Control** the control group.
- Customize the number of permutations for the multi- origin rank product test (at least 100). This number depends on the size of the user dataset (more permutations for smaller dataset), the expected results accuracy (more permutations for more accurate results), and computer performances (more permutation requires higher performance computers)
- Click "**OK**" to start the analysis.

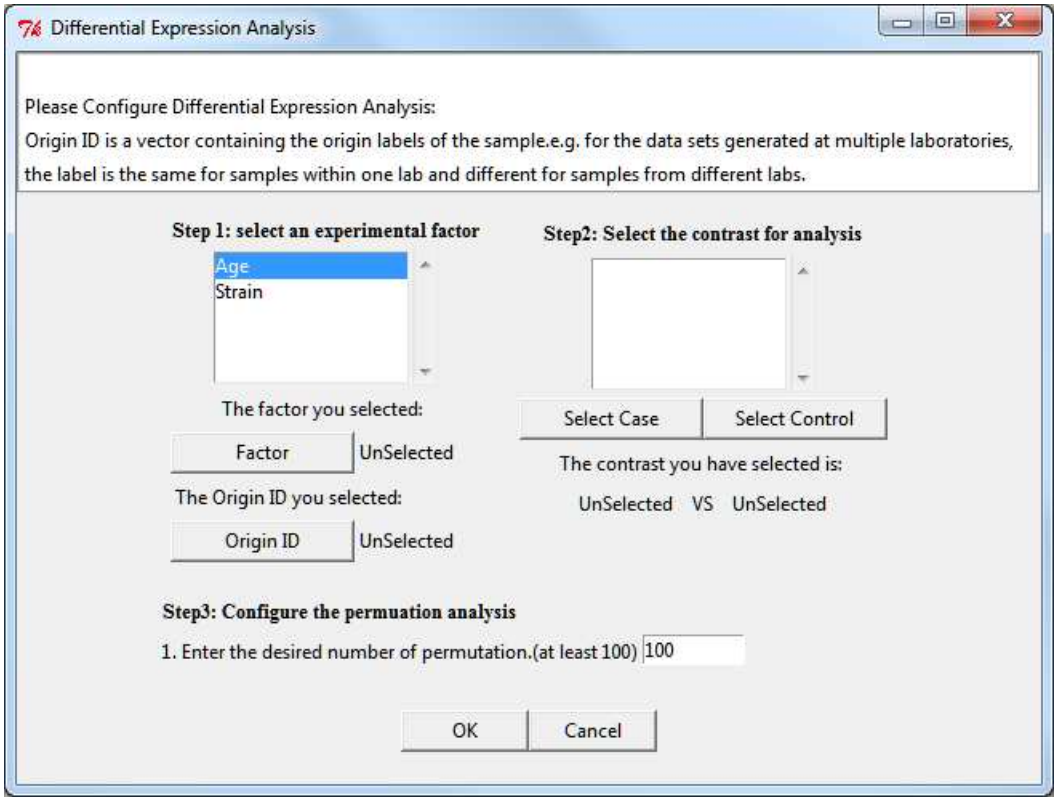


Fig. 8-9 Differential Expression Analysis > RankProduct_Model > RankProd_MultiOrigin

8.4 TIME COURSE DATA ANALYSIS

Microarray Я US implements Bioconductor's maSigPro (Conesa, Nueda et al. 2006) for the analyses of time course microarray data. It builds a model with two factors: group factor (discrete) and time (continuous). The model is assumed to be in the second order of time. For details, please refer to **Appendix 3: List of the implemented key methods**. Select **Differential Expression Analysis > maSigPro > maSigPro_TimeCourse** to invoke the configuration dialogue (Fig. 8-10).

- Select the experimental factor from the available factor listed in **Step 1: Select an experimental factor** box, and click on the **Factor** button to choose it.
- Select the time factor from the available factor listed in **Step 1: Select an experimental factor** box, and click on the **Time** button to choose it.
- Select the replicate factor from the available factor listed in **Step 1: Select an experimental factor** box, and click on the **Replicate** button to choose it.
- Select the contrasts from the available groups listed in the **Step 2: Select the contrast for analysis** box, and click on **Select Case** to choose the case group and **Select Control** the control group.
- Customize the False Discovery Rate (FDR) p-value cutoff for finding significantly altered genes and the significance level for finding significant differences.
- Click "**OK**" to start the analysis.

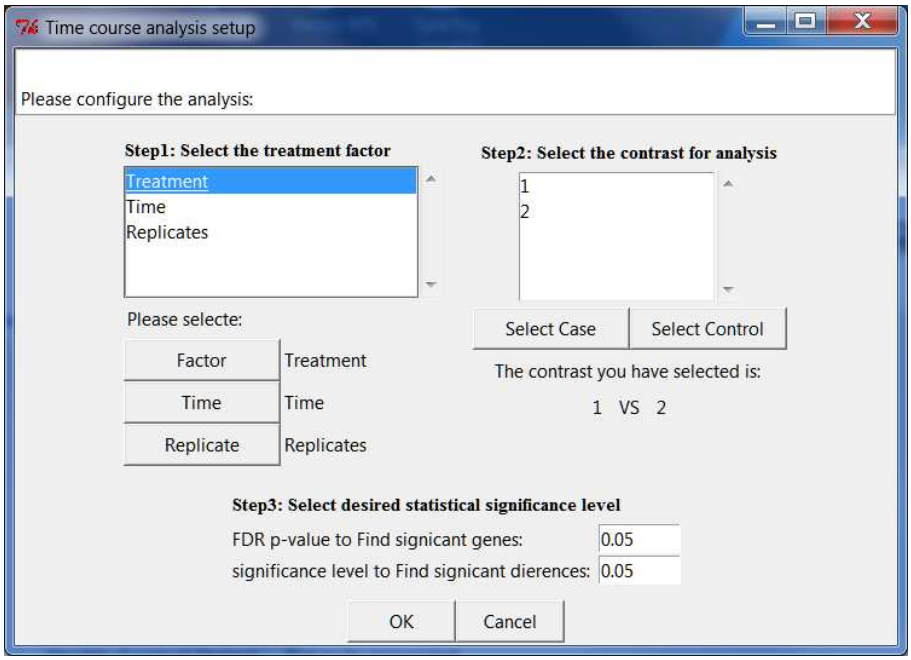


Fig. 8-10 Differential Expression Analysis > maSigPro > maSigPro_TimeCourse

CHAPTER 9. POWER ANALYSIS

Power analysis of a microarray experiment is used to decide 1) the sample size for accurate and reliable statistical judgments, 2) given the sample size, the detection efficiency of the statistical test, and 3) given the sample size, the detection efficiency of fold changes. Accordingly, three types of power analysis are available with the implementation of Biocoductor's ssize package: sample size, power and fold-change. For details, please refer to **Appendix 3: List of the implemented key methods**. To invoke the analysis, select **Power Analysis** from the top menu bar.

Regardless of the type of power analysis to be performed, the configuration dialogue (Fig. 9-1 Power Analysis > Fold Change Fig. 9-1) requires a treatment factor, a control in the selected treatment factor, and parameters for the power analysis. Available treatment factors will be listed in the **Step1: Select the treatment factor** box. Once selected, available sample labels for that treatment factor will be listed in the **Step2: Select the control for analysis** box. Specify proper power, sample size (based on your experimental design), family-wise type 1 error rate for the power analysis and click **OK** to perform the analysis.

Results will be stored as PDF files in the **Output/PowerAnalysis** folder within your project folder.

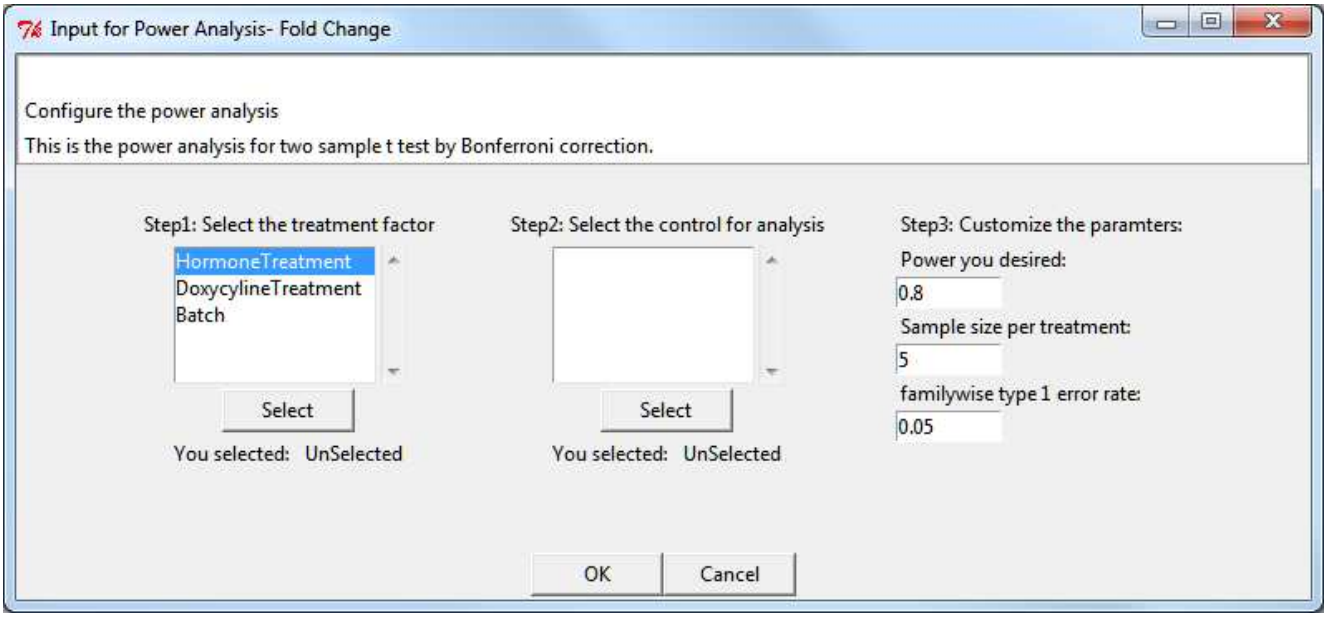


Fig. 9-1 Power Analysis > Fold Change

CHAPTER 10. RESULTS OUTPUT

Results Output has two basic and one advanced functions. The first basic function is to output full table of analysis results and generate differentially expressed gene (DEG) lists. To use this function, users must complete all preceding steps in the workflow. The second basic function is to visualize DEG list(s) via heatmaps or Venn Diagrams. This function can be applied to gene lists generated either by Microarray Я US or by other microarray analysis software.

The advanced function of the Results Output Utility, called Gene List Output Utility, is designed to format microarray analysis results into files that can be directly imported into over 20 widely used tools for functional analysis of microarray results. This function significantly facilitates microarray functional analysis by drastically cutting down the time and efforts required for generating input file of required formats. This function can be applied to both Microarray Я US and other microarray analysis software results. Microarray analysis results generated in other software, however, would require additional simple format changes before being processed.

10.1 GENERATE GENE LISTS

Generate gene list function outputs analysis results into tab delimited (.txt) files. Select **Results Output > Generate Gene List** to invoke the configuration dialogue (Fig. 10-1).

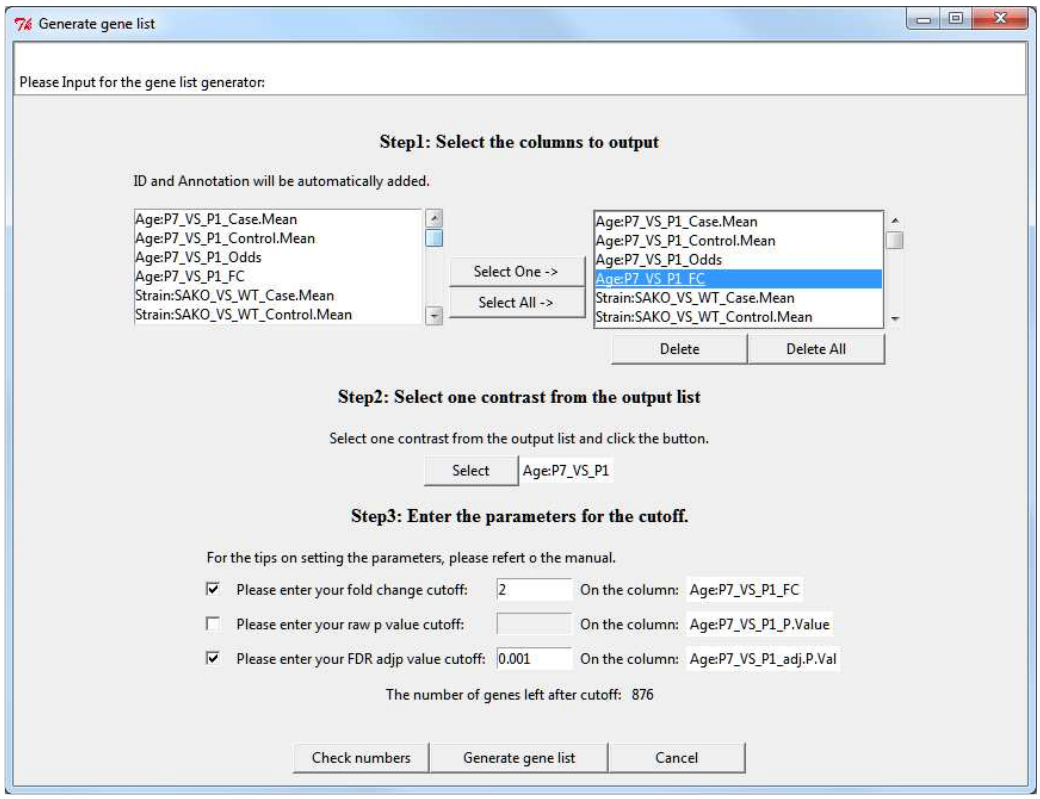


Fig. 10-1 Output Results > Generate Gene List

STEP 1: SELECT COLUMNS TO BE INCLUDED IN THE OUTPUT FILES.

Available data columns will be automatically included in the left text box. Click on individual column names and **Select One ->** to export specific columns, or click on **Select All ->** to export all columns. Probe IDs and all available annotations will be automatically included in the resulting file.

STEP 2: SELECT ONE CONTRAST FROM THE OUTPUT LIST

To generate a DEG list, a contrast (expression fold changes) needs to be specified. To do so, click on any column name with the desired contrast factor names from the output columns list (right text box) and **Select** to confirm.

STEP 3: ENTER THE PARAMETERS FOR THE CUTOFF

Input desired fold-change, raw p-value and FDR adjusted p-value (adj.p) cutoffs to generate DEGs. Click the **Check numbers** button on the bottom to see the number of DEGs selected by the specified cutoff(s). Adjust the cutoff values and click on **Generate gene list** to export.

STEP 4: OUTPUT THE DEG AND THE FULL EXPRESSION FILES

Two results files will be produced, one is the DEG list file selected by specified cutoffs (.DEG.txt), the other a no-cut thus complete gene list file (.DEGcomplete.txt). Both files include all the values and annotations specified in the first step. The gene list output dialogue (Fig. 10-2) allows users to specify the names of the two output files. To help users track the key methods used, the default file names automatically include the employed variation model, the experimental factors, and the selected contrast. Shorten the output file names if desired.

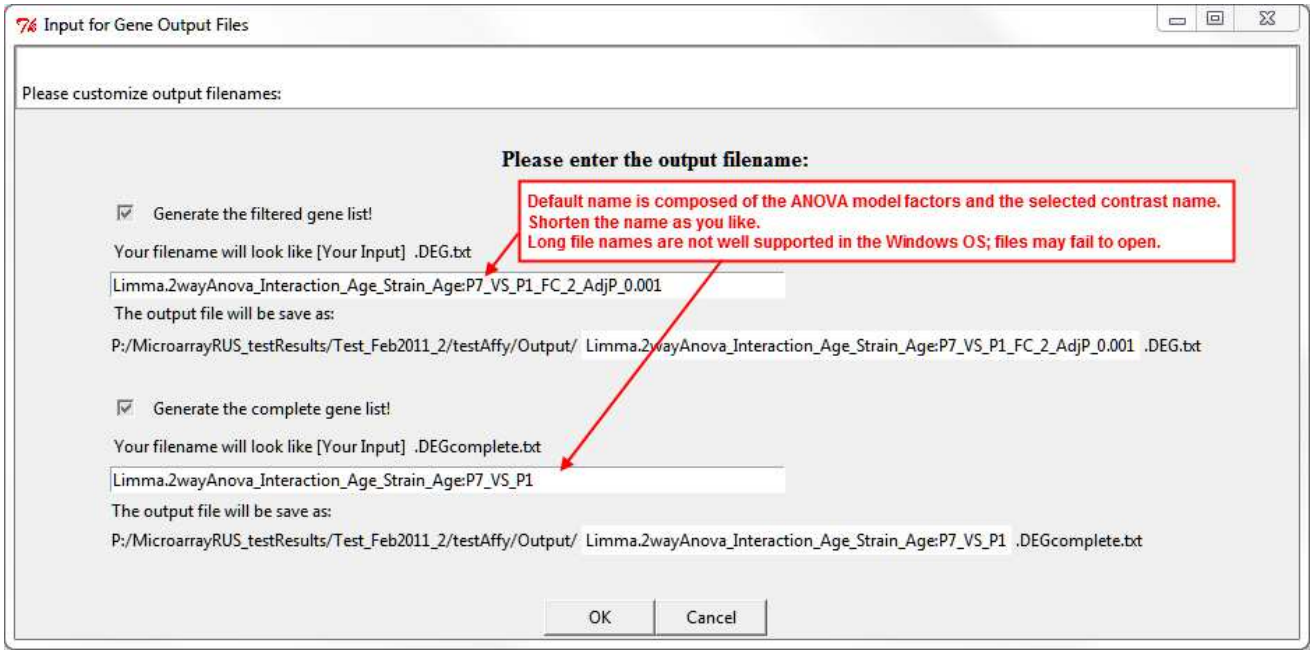


Fig. 10-2 Output Results > Generate Gene List > Name output files

10.2 INSPECT GENE LISTS

Output DEG list and full gene list files include probe ID, all available annotations, and the data columns specified in step1. You may inspect the outputted DEG list file by selecting **Results Output > Inspect Gene List** (Fig. 10-3).

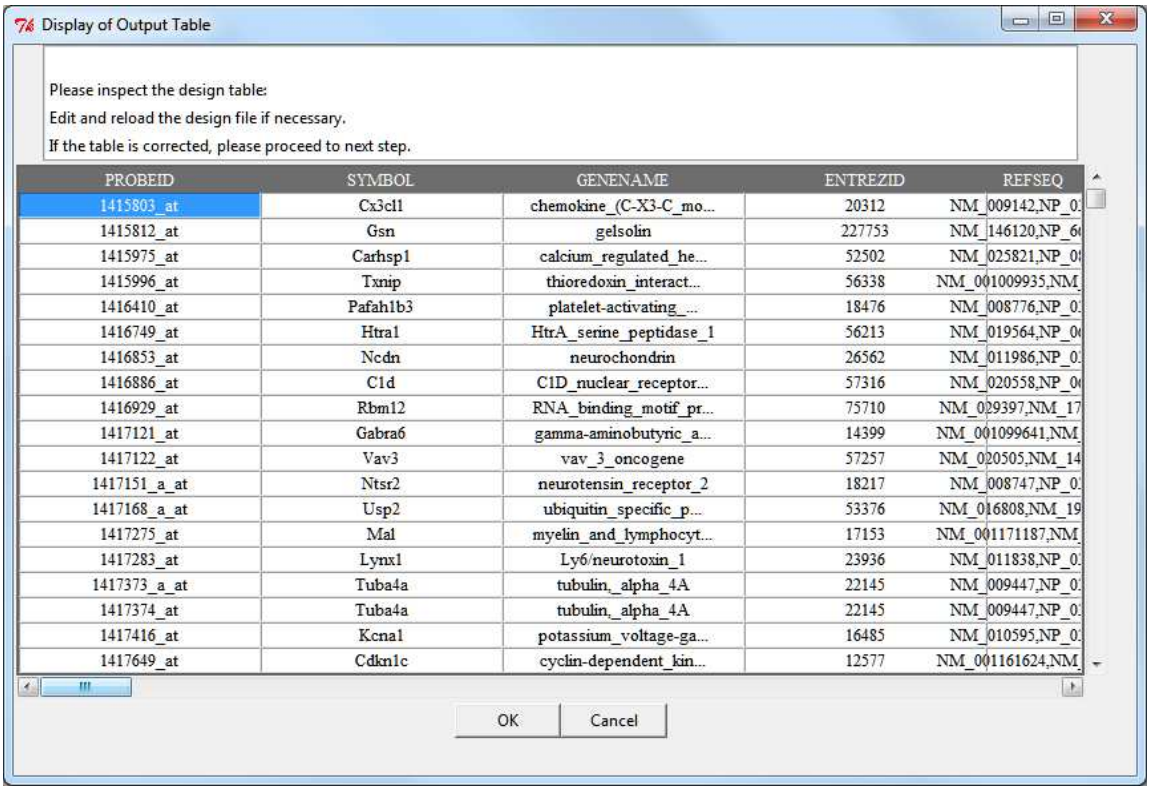


Fig. 10-3 Output Results > Inspect Gene List

10.3 HEATMAP OF DIFFERENTIALLY EXPRESSED GENES

Heatmaps of DEG list allow direct visualization of the analysis results. The **Draw heatmap based on DEG list** function in Microarray Я US can be applied to both Microarray Я US analysis results and external analysis results. To invoke the function, select **Results Output > Draw heatmap based on DEG list** (Fig. 10-4).

- To generate a heatmap, two types of files are required: a DEG list file (*.DEG.txt) and a preprocessed expression file (*.Prep.txt). If using Microarray Я US analysis results, both can be found in the **Output** folder. If using external analysis results, follow **Appendix 7: Notes on Folders and Files** and **Appendix 8: Tutorial for Preparing Partek Genomics Suite (Partek GS) Analysis Results to Use the Gene List Output Utility** to prepare both types of files.

- Heatmap results will be saved as a PDF files in the **Output/Heatmap** folder within your project folder.

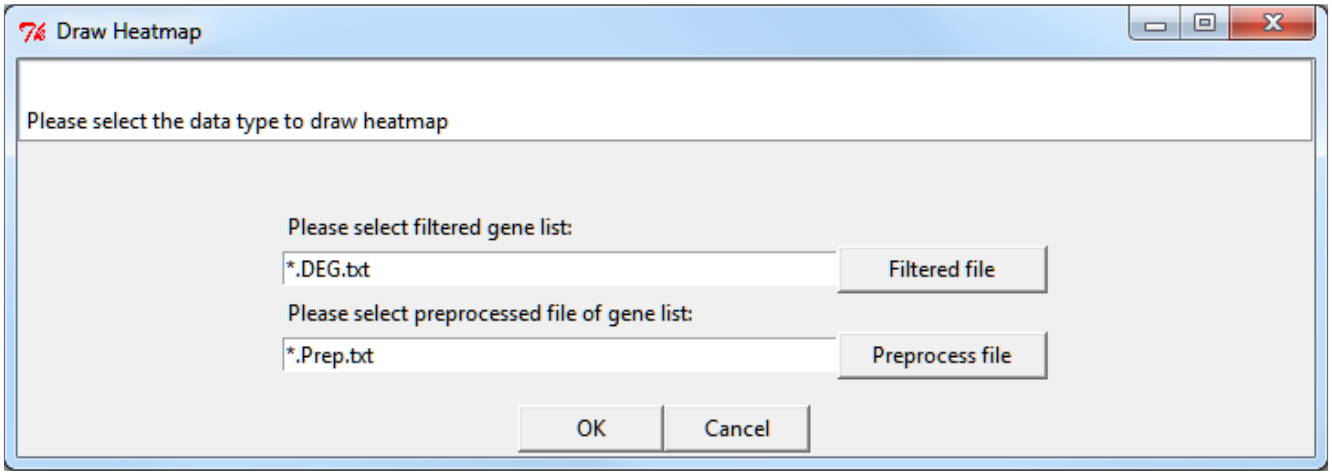


Fig. 10-4 Output Results > Generate heatmap based on DEG list

10.4 VENN DIAGRAM

Venn diagram allows visualization of comparisons between different DEG lists. The **Draw Venn Diagram** function in Microarray Я US can be applied to gene lists generated in Microarray Я US or other microarray analysis software. To invoke the function, select **Results Output > Draw Venn Diagram** (Fig. 10-5).

- To generate a Venn Diagram, two or three DEG list files (*.DEG.txt) are required. If using Microarray Я US analysis results, both can be found in the **Output** folder. If using external analysis results, follow **Appendix 7: Notes on Folders and Files** and **Appendix 8: Tutorial for Preparing Partek Genomics Suite (Partek GS) Analysis Results to Use the Gene List Output Utility** to prepare both types of files.

- The results will be saved as a PDF file in the **Output/Venn** folder and a txt file of the overlapping gene lists in the **Output** folder in your project folder.

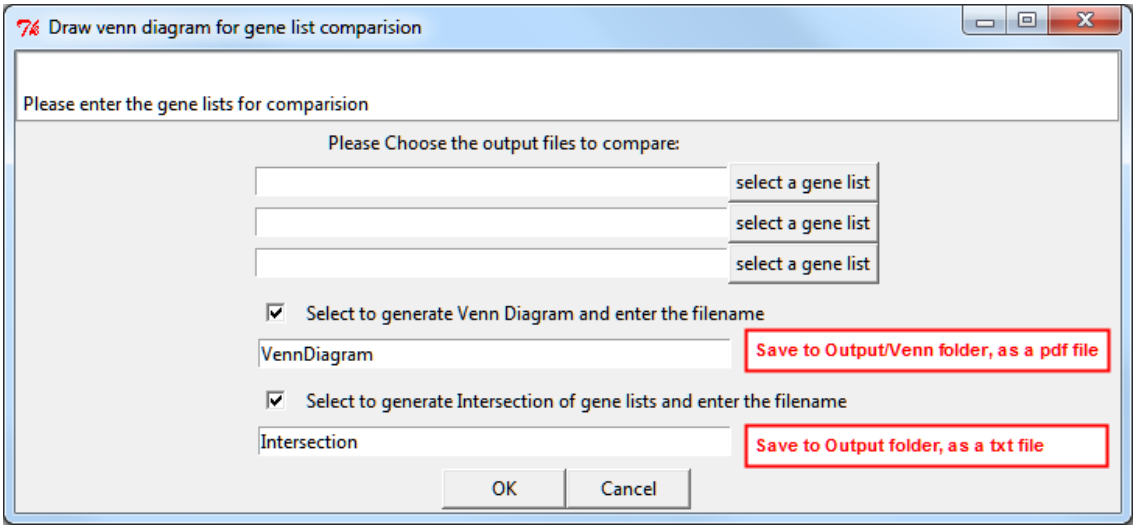


Fig. 10-5 Output Results > Draw Venn Diagram

Notes for Draw Venn Diagram

*The intersections between gene list files are generated using the annotations in the **FIRST column, the column header of which must be named as PROBEID**. When using the Draw Venn Diagram function for external files or Microarray Я US DEG files generated with different CDFs, make sure all files MUST have matching annotations for the PROBEID column (note that Dai's CDF uses similar probe IDs as Affy CDF, but they are actually two completely different systems). When necessary, manually re-arrange column order and edit column header in Excel to use this function.*

10.5 GENE LIST OUTPUT UTILITY

The **Gene List Output Utility** function in Microarray Я US exports DEG lists into input files for over 20 common used gene function analysis software with corresponding formats. This function can be applied to results from Microarray Я US or other microarray analysis software. To invoke the function, select **Results Output > gene list output utility** (Fig. 10-6).

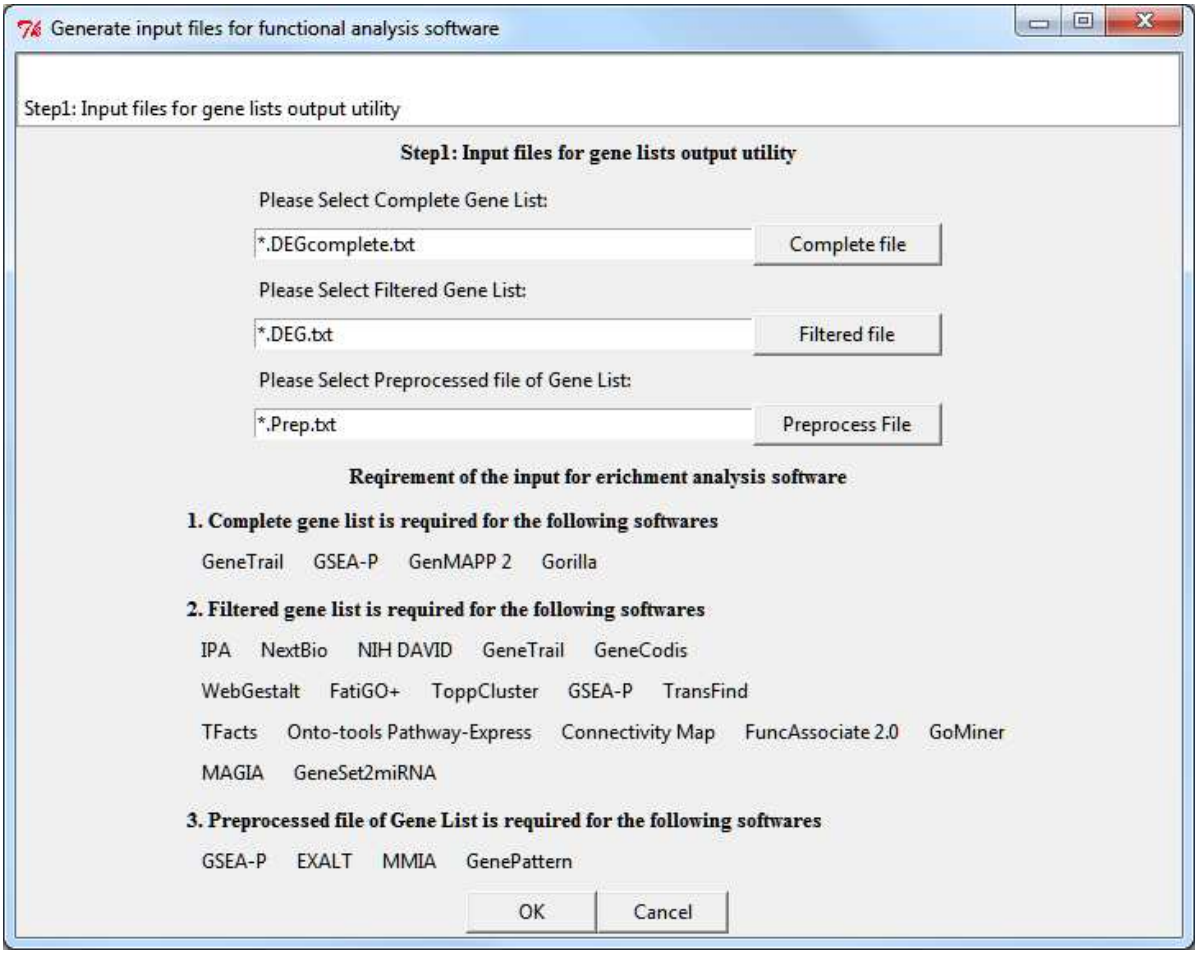


Fig. 10-6 Output Results > Gene List Output Utility

- Depending on the specific format requirements of a given functional analysis software, up to three types of microarray analysis files may be required to generate the corresponding input files: the complete gene list file (*.DEGcomplete.txt), the DEG list file (*.DEG.txt) and the preprocessed expression file (*.Prep.txt). For Microarray Я US analysis results, they can be found in the **Output** folder. If using external analysis results, follow **Appendix 7: Notes on Folders and Files** and **Appendix 8: Tutorial for Preparing Partek Genomics Suite (Partek GS) Analysis Results to Use the Gene List Output Utility** to prepare both types of files. Click **OK** to go to the function analysis software selection window (Fig. 10-7).

- On the function analysis software selection window, select desired enrichment analysis software (Fig. 10-7).

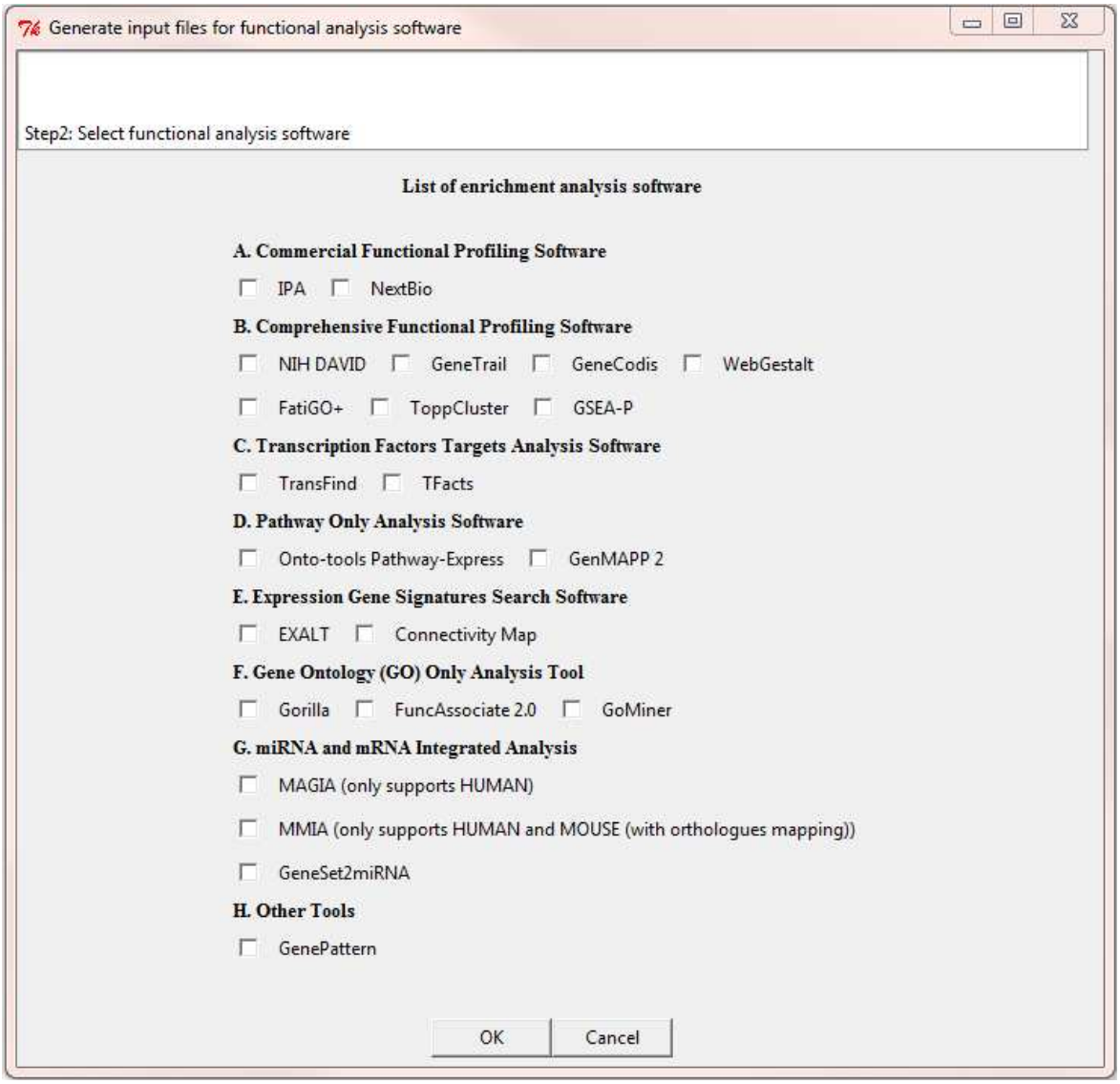


Fig. 10-7 Output Results > Gene List Output Utility> function analysis software selection window

- If **GenePattern** and/or **GSEA-P** software are selected, one additional dialogue window will pop up asking for experimental design file (Fig. 10-8). Click **Select Design File** button to select the file (Refer to **Appendix 7: Notes on Folders and Files**) and **Load File** to import. Available experimental factors will be automatically listed in the **Select the treatment factor** box. Click on the factor to be analyzed and the **Select** button to confirm. Click **OK** to proceed.

- The output files will be named as Factor.Softwarename.txt and saved in the **Output/Utility** folder within your project folder. These files can be directed imported into the corresponding functional analysis software.

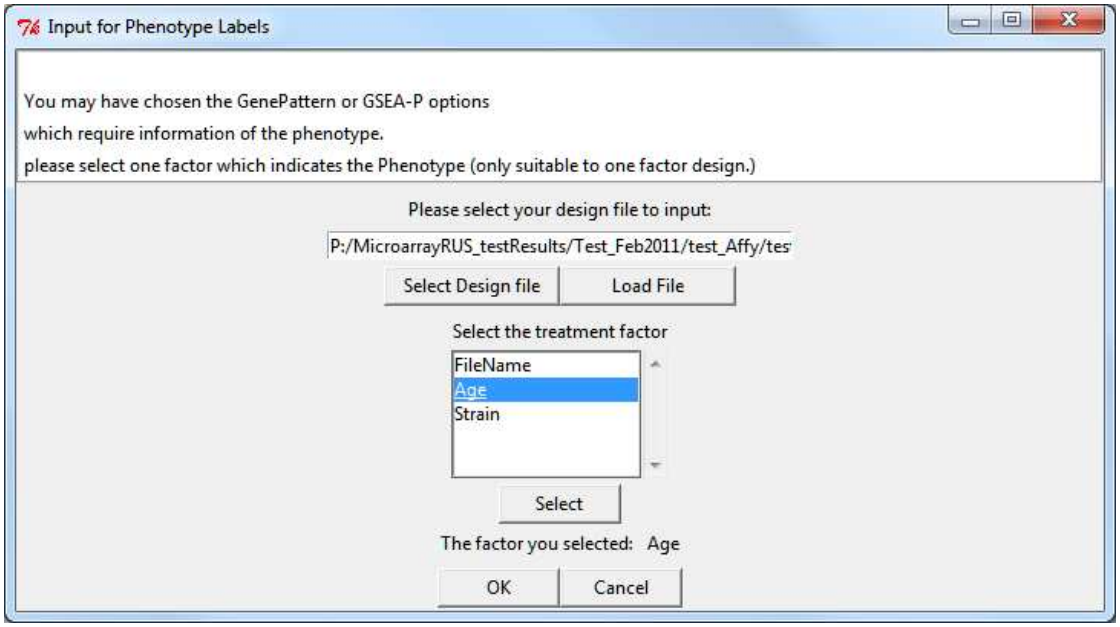


Fig. 10-8 Output Results > Gene List Output Utility > Select Design File for GenePattern and/or GSEA-P

TERMS OF USE

APPENDIX

APPENDIX 1. LIST OF THE SUPPORTED MICROARRAY DATA TYPES

Affymetrix Intensity Data (.CEL files)	
Human	HG-U133A; HG-U133A_2; HG-U133B; HG-U133_Plus_2; HG_U95A; HG_U95Av2; HG_Focus
Mouse	MG_U74Av2; MG_U74Bv2; MG_U74Cv2; Mouse430_2; Mouse430A_2; MOE430A; MOE430B
Rat	RG_U34A; Rat230_2; RAE230A; RAE230B
Illumina Expression Beadchips (BeadStudio outputs)	
Human (version 1 to 3)	HumanWG-6; HumanRef-8; HumanHT-12
Mouse (version 1 to 2)	MouseWG-6; MouseRef-8
Rat (version 1)	RatRef-12

APPENDIX 2. LIST OF THE KEY BIOCONDUCTOR PACKAGES IMPLEMENTED

PUBLIC DATA ACCESS PACKAGES

- **GEOquery** ---- Get data from NCBI Gene Expression Omnibus (GEO) (Sean and Meltzer 2007)
- **Geometadb** ---- A compilation of metadata from NCBI GEO (Zhu, Davis et al. 2008)
- **ArrayExpress** ---- Access the ArrayExpress Microarray Database at EBI and build Bioconductor data structures (Kauffmann, Rayner et al. 2009)

PREPROCEESING AND NORMALIZATION PACKAGES

- **affy** ---- Methods for processing Affymetrix oligonucleotide arrays (Gautier, Cope et al. 2004)
- **lumi** ---- Methods for processing Illumina BeadArrays (Du, Kibbe et al. 2007; Du, Kibbe et al. 2008; Lin, Du et al. 2008)
- **vsn** ---- Variance stabilization and calibration for microarray data (Huber, von Heydebreck et al. 2002)
- **gcrma** ---- Background adjustment method using sequence information (Wu, Irizarry et al. 2002)

QUALITY CONTROL PACKAGES

- **arrayQualityMetrics** ---- Quality metrics on ExpressionSets (Kauffmann, Gentleman et al. 2009)
- **affyQCReport** ---- QC Report Generation for affyBatch objects (Parman, Halling et al. 2010)

DIFFERENTIALLY EXPRESSED GENE DETECTION PAKCAKGES

- **limma** ---- Linear models for microarray data (Smyth 2004)
- **siggenes** ---- SAM and Efron's empirical Bayes approaches (Schwender 2009)
- **RankProd** ---- Rank product method for identifying differentially expressed genes with application in meta-analysis (Hong, Wittner et al. 2009)
- **maSigPro** ---- Microarray significant gene expression profile – find differences in time course data (Conesa, Nueda et al. 2006)

POWER ANALYSIS PACKAGES

- **ssize** ---- Estimate microarray sample size (Warnes, Liu et al. 2009)

APPENDIX 3. LIST OF THE IMPLEMENTED KEY METHODS

AFFYMETRIX PREPROCESSING AND NORMALIZAITON METHODS

- **RMA** (Robust Multi-Array Average Method) (Irizarry, Hobbs et al. 2003)
- **gcRMA** (RMA using sequence information) (Wu, Irizarry et al. 2004; Wu and Irizarry 2005)
- **MAS5** (MAS 5.0 Method) (Affymetrix 2002)
- **dChip** (Li and Wong) (Li and Wong 2001)
- **Advanced** (User can choose their desired method for each preprocessing step)

For details, please refer to the vignette of the **affy package** (Gautier, Cope et al. 2004)

ILLUMINA PREPROCESSING AND NORMALIZATION METHODS

- **Background Correction**: None; Background Adjust; Force Positive; Background Adjust by using method implemented in the **affy package** (Gautier, Cope et al. 2004)
- **Normalization**: Quantile Normalization; Robust Spline Normalization; Simple Scaling Normalization; Scale by LOESS; Variance stabilization and calibration; Rank Invariant Normalization
- **Variance stabilizing transformation**: Variance-Stabilizing transformation; Log2 transformation; Cubic Root transformation

For details, please refer to the vignette of the **lumi package** (Du, Kibbe et al. 2007; Du, Kibbe et al. 2008; Lin, Du et al. 2008).

DIFFERENTIAL EXPRESSION ANALYSIS METHODS

LINEAR MODEL (LIMMA PACKAGE)

Linear models for one factor, two factor, random block designs, and multi-factor with block designs (Smyth 2004) are implemented in the Microarray Я US. The differentially expressed genes are detected based on a Bayesian moderated t-test. This model provides reliable results even for experiments with small sample sizes.

SAM (SIGGENES PACKAGE)

The Significance Analysis of Microarrays (SAM) method (Tusher, Tibshirani et al. 2001) is a permutation test based on either a modified t-statistics or a Wilcoxon rank statistics. This method can be used for analyzing both paired and unpaired two-class experimental designs.

RANK PRODUCT TEST (RANKPROD PACKAGE)

Rank Product test (Breitling, Armengaud et al. 2004) is a non-parametric statistical method based on the fold change ranks of each gene. This method can be used for two-class experimental designs analysis as well as meta-analysis.

TIME COURSE DATA ANALYSIS (MASIGPRO PACKAGE)

The method implemented in maSigPro is a regression based method (Conesa, Nueda et al. 2006). The underlying model consists of two factors: the group factor (discrete) and the time (continuous) variant. In Microarray Я US, we assume that the model is in the second order of time. The significant genes can be extracted based on user-specified contrasts.

APPENDIX 4. LIST OF THE IMPLEMENTED CUSTOM CDF AND ANNOTATIONS

AFFYMETRIX CUSTOM CDF BY DAI ET AL., (2005) VERSION 13 (MARCH,2011)

- **Web Site** describing of custom CDF (Dai, Wang et al. 2005)

AFFYMETRIX CUSTOM CDF BY ALBERTO RISUENO ET AL., (2010)

- **Web Site** describing of custom CDF (Prieto, Risueno et al. 2008; Risueno, Fontanillo et al. 2010)

ILLUMINA REANNOTATION BY BARBOSA-MORAIS ET AL., (2009)

- **Web Site** describing of Illumina BeadArray probe reannotation (Barbosa-Morais, Dunning et al. 2010)

Re-annotated Beadchip Types	
Human	Human WG-6 version 1, 2, 3 Human Ref-8 version 1, 2, 3 Human DASL
Mouse	Mouse WG-6 version 1, 1.1, 2 Mouse Ref-8 version 1, 1.1, 2
Rat	Rat Ref-12 version 1

ILLUMINA RE-ANNOTATION BY DU ET AL., (2007)

- This is the annotation implemented in **lumi package**. For details, please refer to the vignette of **lumi package**. (Du, Kibbe et al. 2007)

Re-annotated Beadchip Types	
Human	Human WG-6 version 1, 2, 3 Human Ref-8 version 1, 2, 3 Human HT12 version 2, 3
Mouse	Mouse WG-6 version 1, 2 Mouse Ref-8 version 1, 2
Rat	Rat Ref-12 version 1

APPENDIX 5. LIST OF THE SUPPORTED FUNCTIONAL ANALYSIS SOFTWARE

5.A COMMERCIAL FUNCTIONAL PROFILING SOFTWARE

5.A.1 INGENUITY PATHWAY ANALYSIS (IPA)

Web Site:	**http://www.ingenuity.com/**
Required ID:	All major IDs accepted, allows multiple ID columns
Required data type:	DEG list with Probe ID, Gene Symbol, FC, p, FDR-P
Supported organisms:	Human, mouse, rat; ortholog gene mapping for other major model organisms
Required file format:	Tab delimitated .txt file
Output File Name:	*.IPA.txt

5.A.2. NEXTBIO

Web Site:	**http://www.nextbio.com/**
Required ID:	All major IDs accepted, allows multiple ID columns
Required data type:	DEG list—Probeset ID, Gene Symbol, along with FC and p and FDR-adjusted p
Supported organisms:	Many major model organisms
Required file format:	Tab delimitated .txt file
Output File Name:	*.NextBio.txt
Functional analysis type:	SEA (multiple lists allowed)
Content type:	Mixed (Human curated and computational predicted)
Major functional analysis categories:	Pathways, GO, TF targets, miRNA targets, disease, protein domains, SNP, chromosomal locations, literatures

5.B COMPREHENSIVE FUNCTIONAL PROFILING SOFTWARE

5.B.1 NIH DAVID (HUANG DA, SHERMAN ET AL. 2009; HUANG DA, SHERMAN ET AL. 2009)

Web Site:	**http://david.abcc.ncifcrf.gov/**
Required ID:	Probe set ID (for all standard CDF Affymetrix and Illumina arrays)
Required data type:	DEG list-Probe set ID only
Supported organisms:	Many major model organisms
Required file format:	Tab delimitated .txt file
Output File Name:	*.DAVID.txt
Required ID:	Gene symbol (for all custom CDF Affymetrix arrays)
Required data type:	(1) DEG list-with Gene Symbol only (2) Background list
Supported organisms:	Many major model organisms

Required file format:	Tab delimitated .txt file
Output File Name:	*.CDF-DAVID.txt
Functional analysis type:	SEA and MEA
Content type:	Mixed (Human curated and computational predicted)
Major functional analysis categories:	Pathways (multiple database), GO, TF targets, miRNA targets, disease, protein domains, protein-protein interactions (multiple databases), GWAS, chromosomal locations

5.B.2 GENETRAIL—ADVANCED GENE SET ENRICHMENT ANALYSIS (BACKES, KELLER ET AL. 2007)

Web Site:	**http://genetrail.bioinf.uni-sb.de/**
Required ID:	Gene Symbol
Required data type:	(1) DEG list - symbol only or the complete variance analyzed gene list-with gene symbol only, ranked by p (GSEA mode) (2) Reference list (for data processed with customized CDFs only)
Supported organisms:	Many major model organisms
Required file format:	Tab delimitated .txt file
Output File Name:	*.Genetrail-SEA.txt or *.Genetrail-GSEA.txt (GSEA mode)
Functional analysis type:	SEA and GSEA
Content type:	Mixed (Human curated and computational predicted)
Major functional analysis categories:	Pathways, GO, TF targets, miRNA targets, disease, protein domains, SNP, chromosomal locations

5.B.3 GENECODIS: INTERPRETING GENE LISTS THROUGH ENRICHMENT ANALYSIS AND INTEGRATION OF DIVERSE BIOLOGICAL INFORMATION (NOGALES-CADENAS, CARMONA-SAEZ ET AL. 2009)

Web Site:	**http://genecodis.dacya.ucm.es/**
Required ID:	Gene Symbol
Required data type:	(1) DEG list-Gene Symbol only (2) Reference list (optional, only for data processed with a customized CDF)
Supported organisms:	Many major model organisms
Required file format:	Tab delimitated txt file.
Output File Name:	*.GeneCodis.txt
Functional analysis type:	SEA and MEA
Content type:	Mixed (Human curated and computational predicted)
Major functional analysis categories:	Pathways, GO (different levels and GOSlim), TF targets, miRNA targets, protein motifs

5.B.4 WEBGESTALT: AN INTEGRATED SYSTEM FOR EXPLORING GENE SETS IN VARIOUS BIOLOGICAL CONTEXTS (ZHANG, KIROV ET AL. 2005)

Web Site:	**http://bioinfo.vanderbilt.edu/wg_gsat/**
Required ID:	Gene Symbol
Required data type:	DEG list—Gene Symbol with FC

Supported organisms:	Many major model organisms
Required file format:	Tab delimitated txt file.
Output File Name:	*_WebGestalt.txt
Functional analysis type:	SEA
Content type:	Mixed (Human curated and computational predicted)
Major functional analysis categories:	Pathways (multiple databases), GO, TF targets, miRNA targets, protein-protein interaction, chromosomal locations
Note:	The GO analysis module is from the popular GOTM (GOTree Machine 2004, c=306)

5.B.5 FATIGO +: A FUNCTIONAL PROFILING TOOL FOR GENOMIC DATA. INTEGRATION OF FUNCTIONAL ANNOTATION, REGULATORY MOTIFS AND INTERACTION DATA WITH MICROARRAY EXPERIMENTS (AL-SHAHROUR, MINGUEZ ET AL. 2007)

Web Site:	**http://babelomics.bioinfo.cipf.es/functional.html**
Required ID:	Gene Symbol
Required data type:	DEG list-Gene symbol only
Supported organisms:	Many major model organisms
Required file format:	Tab delimitated .txt file
Output File Name:	*.FatiGO.txt
Web Site:	http://babelomics.bioinfo.cipf.es/functional.html
Functional analysis type:	SEA
Content type:	Mixed (Human curated and computational predicted)
Major functional analysis categories:	Pathways, GO, GOSlim, TF targets, Regulatory sequences; miRNA targets, protein domains;
Note:	Allows customized level setting for GO analysis.

5.B.6 TOPPCLUSTER: A MULTIPLE GENE LIST FEATURE ANALYZER FOR COMPARATIVE ENRICHMENT CLUSTERING AND NETWORK-BASED DISSECTION OF BIOLOGICAL SYSTEMS (KAIMAL, BARDES ET AL. 2010)

Web Site:	**http://toppcluster.cchmc.org/**
Required ID:	Gene Symbol
Required data type:	DEG list-symbol only
Supported organisms:	**mainly human, mouse and rat also workable**
Required file format:	Tab delimitated .txt file
Output File Name:	*.ToppCluster.txt
Functional analysis type:	SEA
Content type:	Mixed (Human curated and computational predicted)
Major functional analysis categories:	Pathways, GO, TF targets, miRNA targets, disease, protein domains; protein-protein interaction, drugs, human/mouse phenotypes; co-expression gene sets; chromosomal locations, literatures
Note:	Allows multiple gene lists comparison

5.B.7 GSEA-P: A DESKTOP APPLICATION FOR GENE SET ENRICHMENT ANALYSIS (SUBRAMANIAN, KUEHN ET AL. 2007)

Web Site:	**http://www.broad.mit.edu/GSEA**
Required ID:	All major IDs accepted, allows multiple ID columns
Required data type:	(1) .GCT file of preprocessed data—Gene Symbol, natural scale intensity data for each sample (2) .CLS file of phenotype labels (3). RNK file of a variance analyzed completed gene list---Gene Symbol only, pre-ranked based on p value
Supported organisms:	Many major model organisms
Required file format:	Tab delimitated files saved in .gct, .cls, .rnk format
Output File Name:	*.GSEA.gct *.GSEA.cls *.GSEA.rnk
Functional analysis type:	GSEA
Content type:	Mixed (Human curated and computational predicted)
Major functional analysis categories:	Pathways, GO, TF targets, miRNA targets, various expression gene sets; chromosomal locations
Note:	GSEA requires either one .rnk file OR both .gct and .cls files for the analysis

5.C TRANSCRIPTION FACTORS TARGETS ANALYSIS SOFTWARE

5.C.1 TRANSFIND—PREDICTING TRANSCRIPTIONAL REGULATORS FOR GENE SETS (KIELBASA, KLEIN ET AL. 2010)

Web Site:	**http://transfind.sys-bio.net/**
Required ID:	Gene Symbol
Required data type:	(1) DEG list—Gene Symbol only (2) Reference list (optional, only for data processed with a customized CDF)
Supported organisms:	Many major model organisms
Required file format:	Tab delimitated .txt file
Output File Name:	*.TransFind.txt
Functional analysis type:	SEA
Content type:	Mixed (Human curated and computational predicted)
Major functional analysis categories:	Transcription factors with conserved binding motif

5.C.2 TFACTS—TRANSCRIPTION FACTOR REGULATION CAN BE ACCURATELY PREDICTED FROM THE PRESENCE OF TARGET GENE SIGNATURES IN MICROARRAY GENE EXPRESSION DATA (ESSAGHIR, TOFFALINI ET AL. 2010)

Web Site:	**http://www.tfacts.org/**
Required ID:	Gene Symbol

Required data type:	(1) Up-regulated DEG list-symbol only (2) Down-regulated DEG list-symbol only
Supported organisms:	Primarily human, but also mouse/rat human ortholog gene
Required file format:	Tab delimitated txt file.
Output File Name:	*.UP_TFactS.txt; *.DOWN_TFactS.txt
Functional analysis type:	SEA
Content type:	Human curated
Major functional analysis categories:	Transcription factors regulation (sign-sensitive)

5.D PATHWAY ONLY ANALYSIS SOFTWARE

5.D.1 ONTO-TOOLS PATHWAY-EXPRESS (DRAGHICI, KHATRI ET AL. 2007)

Web Site:	**http://vortex.cs.wayne.edu/projects.htm**
Required ID:	Gene Symbol
Required data type:	(1) DEG list—Gene Symbol with FC (2) Reference list (optional, only for data processed with a customized CDF)
Supported organisms:	Many major model organisms
Required file format:	Tab delimitated txt file.
Output File Name:	*.Onto-PE.txt
Functional analysis type:	SEA
Content type:	Human curated
Major functional analysis categories:	Pathways with impact factors (calculated based on the gene expression directions and the topography of a pathway).

5.D.2 GENMAPP 2 (SALOMONIS, HANSPERS ET AL. 2007)

Web Site:	**http://www.genmapp.org/**
Required ID:	Ensembl Symbol
Required data type:	(1) The complete variance analyzed gene list with proper ID and data (2) Reference list (optional, only for data processed with a customized CDF)
Supported organisms:	Many major model organisms
Required file format:	Tab delimitated txt file.
Output File Name:	*.GenMAPP.txt
Functional analysis type:	SEA
Content type:	Human curated
Major functional analysis categories:	Pathways

5.E EXPRESSION GENE SIGNATURES SEARCH SOFTWARE

5.E.1 EXALT-- WEB-BASED INTERROGATION OF GENE EXPRESSION SIGNATURES USING EXALT (WU, QIU ET AL. 2009)

Web Site:	**http://seq.mc.vanderbilt.edu/exalt/**
Required ID:	ProbeID and Gene Symbol
Required data type:	Preprocessed data (natural scale) with all samples listed
Required experiment type:	One factor with up to 9 levels, with at least 2 replicates in each group
Supported organisms:	Human, mouse, rat
Required file format:	Tab delimitated .txt file
Output File Name:	*_EXALT.txt
Functional analysis type:	General expression signatures mining
Content type:	Human curated
Major functional analysis categories:	Public expression data signatures

5.E.2 THE CONNECTIVITY MAP: USING GENE-EXPRESSION SIGNATURES TO CONNECT SMALL MOLECULES, GENES, AND DISEASE (LAMB 2007)

Web Site:	**http://www.broadinstitute.org/cmap/**
Required ID:	Affymetrix HG-U133A probe set ID, mapped from Gene Symbol
Required data type:	**(**1) Up-regulated DEG list- Affymetrix HG-U133A probe set ID only (2) Down-regulated DEG list- Affymetrix HG-U133A probe set ID only
Supported organisms:	Primarily human, but also mouse/rat human ortholog genes
Required file format:	Tab delimitated txt file.
Output File Name:	*.UP_CMAP.grp; *.DOWN_CMAP.grp
Note:	Total DEG list should not exceed 1000 genes.
Functional analysis type:	Expression signatures mining
Content type:	Human curated
Major functional analysis categories:	Cell-line drug treatment expression signatures

5.F GENE ONTOLOGY (GO) ONLY ANALYSIS TOOL

5.F.1 GORILLA—A TOOL FOR DISCOVERY AND VISUALIZATION OF ENRICHED GO TERMS IN RANKED GENE LISTS (EDEN, NAVON ET AL. 2009)

Web Site:	**http://cbl-gorilla.cs.technion.ac.il/**
Required ID:	Gene Symbol
Required data type:	The complete variance analyzed gene list-with gene symbol only, ranked by p
Supported organisms:	Many major model organisms
Required file format:	Tab delimitated txt file.
Output File Name:	*.GOrilla.txt

Functional analysis type: SEA and GSEA
Content type: Mixed (Human curated and computational predicted)

Major functional analysis categories: GO

5.F.2 FUNCASSOCIATE 2.0-- NEXT GENERATION SOFTWARE FOR FUNCTIONAL TREND ANALYSIS (BERRIZ, BEAVER ET AL. 2009)

Web Site: http://llama.med.harvard.edu/funcassociate/
Required ID: Gene Symbol
Required data type: DEG list(s)--with Gene Symbol only
Supported organisms: Many major model organisms
Required file format: Tab delimitated txt file.
Output File Name: *.FuncAssociate.txt
Functional analysis type: SEA
Content type: Mixed (Human curated and computational predicted)
Major functional analysis categories: GO
Note: Allows Customized GO Evidence Codes Setting

5.F.3 GOMINER (HIGH-THROUGHPUT)—AN INTEGRATIVE GENE ONTOLOGY TOOL FOR INTERPRETATION OF MULTIPLE-MICROARRAY EXPERIMENTS (ZEEBERG, FENG ET AL. 2003)

Web Site: discover.nci.nih.gov/gominer/htgm.jsp/
Required ID: Gene Symbol
Required data type: DEG list(s)--with gene symbol + signs of up/down regulations)
Supported organisms: Many major model organisms
Required file format: Tab delimitated txt file.
Output File Name: *.GoMiner.txt
Web Site: http://discover.nci.nih.gov/gominer/htgm.jsp
Functional analysis type: SEA
Content type: Mixed (Human curated and computational predicted)
Major functional analysis categories: GO
Note: Allows multiple DEG lists comparison

5.G MIRNA AND MRNA INTEGRATED ANALYSIS

5.G.1 MAGIA—A WEB-BASED TOOL FOR MIRNA AND GENES INTEGRATED ANALYSIS (SALES, COPPE ET AL. 2010)

Web Site: http://gencomp.bio.unipd.it/magia
Required ID: Entrez ID
Required data type: DEG list—Entrez ID, preprocessed natural scale intensities data for each sample
Supported organisms: Human only
Required file format: Tab delimitated txt file.
Output File Name: *.MAGIA.txt

Web site:	http://gencomp.bio.unipd.it/magia/start/
Functional analysis type:	Correlation
Content type:	Mixed (Human curated and computational)
Major functional analysis categories:	miRNA targets prediction, miRNA-mRNA expression correlation analysis
Note:	Sample names and order must be matched between the mRNA and miRNA lists.

5.G.2 MMIA : MIRNA AND MRNA INTEGRATED ANALYSIS (NAM, LI ET AL. 2009)

Web Site:	**http://156.56.93.156/~MMIA/index.html**
Required ID:	Gene Symbol
Required data type:	Preprocessed data—Gene Symbol, preprocessed natural scaled intensities data for each sample
Supported organisms:	Human only
Required file format:	Tab delimitated txt file.
Output File Name:	your_file_name.MMIA.txt
Functional analysis type:	GSEA and Correlation
Content type:	Mixed (Human curated and computational)
Major functional analysis categories:	miRNA targets prediction, miRNA-mRNA expression correlation analysis; TFBS in miRNA promoter; diseases; pathways, GO, cancer gene sets, chromosomal locations
Note:	sample name and order must be matched between the mRNA and miRNA lists.

5.G.3 GENESET2MIRNA (ANTONOV, DIETMANN ET AL. 2009)

Web Site:	**http://mips.helmholtz-muenchen.de/proj/gene2mir/**
Required ID:	Gene Symbol
Required data type:	DEG list—Gene Symbol only
Supported organisms:	Human, mouse, rat
Required file format:	Tab delimitated txt file.
Output File Name:	*.MAGIA.txt
Functional analysis type:	Correlation
Content type:	Mixed (Human curated and computational)
Major functional analysis categories:	miRNA targets prediction, miRNA-mRNA expression correlation analysis
Note:	sample name and order must be matched between the mRNA and miRNA lists.

5.H OTHER TOOLS

5.H.1 GENEPATTERN--USING GENEPATTERN FOR GENE EXPRESSION ANALYSIS (KUEHN, LIBERZON ET AL. 2008)

Web Site:	**www.broadinstitute.org/cancer/software/genepattern/**
Required ID:	All major IDs accepted, allows multiple ID columns
Required data type:	(1) GCT file of preprocessed data—Gene Symbol, natural scale intensity data for each sample (2) CLS file of phenotype labels
Supported organisms:	Many major model organisms
Required file format:	Tab delimitated file saved with .gct or .cls
Output File Name:	*.GenePattern.gct *.GenePattern.cls
Analysis type:	Statistical and visual analysis of microarray data
Note:	Over 100 programs available in GenePattern for a wide spectrum of microarray data analysis and manipulation

APPENDIX 6. EXPORT ILLUMINA GENE EXPRESSION DATA FROM BEADSTUDIO

Microarray Я US supports direct import of Illumina raw data from Illumina BeadStudio (v.1 – v.3) export. The following tutorial describes how to export Illumina gene expression data using Illumina BeadStudio for use in the Microarray Я US.

STEP 1: CONFIGURE GENE EXPRESSION DATA IMPORT.

Import gene expression data using the BeadStudio import wizard and configure analysis details on **Please choose analysis type and parameters** dialogue (Fig.App.1). Although Microarray Я US supports importing normalized data**, non-normalized data is preferred**. Normalization can be done later in Microarray Я US, please refer to Step 4 for details. To import non-normalized data,

- Select **Gene Expression** for Analysis Type
- Select **none** for Normalization, and **uncheck** the **Subtract Background** box

STEP 2: CONFIGURE SAMPLE COLUMNS TO BE INCLUDED IN THE EXPORT FILE.

- Select the "Sample Probe Profile" tab in the **Gene Expression Analysis** window (Fig.App.2).
- Click on the "Column Chooser" icon to specify the columns to export (Fig. App.2).
- In the "Column Chooser" window (Fig. App.3), select **PROBE ID** and data columns in the "Displayed Columns" box, and **AVG Signal**, **BEAD STDERR** in the "Displayed Subcolumns" box. These are the minimum required columns to be included in the sample data for Microarray Я US analysis. Additional columns will be ignored in Microarray Я US.

STEP 3: EXPORT SAMPLE DATA

- Click **OK** to go back to the "Sample Probe Profile" tab in the Gene Expression Analysis window (Fig.App.2).
- Click on "Export displayed data" icon to export sample data (as .txt file) (Fig.App.2).
- Refer to Fig.App.4 for an example of the exported file for Microarray Я US import.

STEP 4 (OPTIONAL): EXPORT BACKGROUND CONTROL PROFILE

In order to perform **Background Adjust Correction** in Microarray Я US, an additional file, the Background Control file, needs to be exported from BeadStudio and imported into Microarray Я US.

- Select the "Control Probe Profile" tab in the Gene Expression Analysis window (Fig.App.5).

- Click on the “Column Chooser” icon to specify the columns to export (Fig.App.5).
- In the “Column Choose” window, select **TargetID**, **ProbeID** and data columns in the “Displayed Columns” box, and **AVG Signal** and **Detection Pval** in the “Displayed Subcolumns” box (Fig.App.6). These are the minimum required columns to be included in the sample data for Microarray Я US analysis. Additional columns are ignored in Microarray Я US.
- Export and name your control profile (.txt file). This will be the input background correction file in Microarray Я US**.**

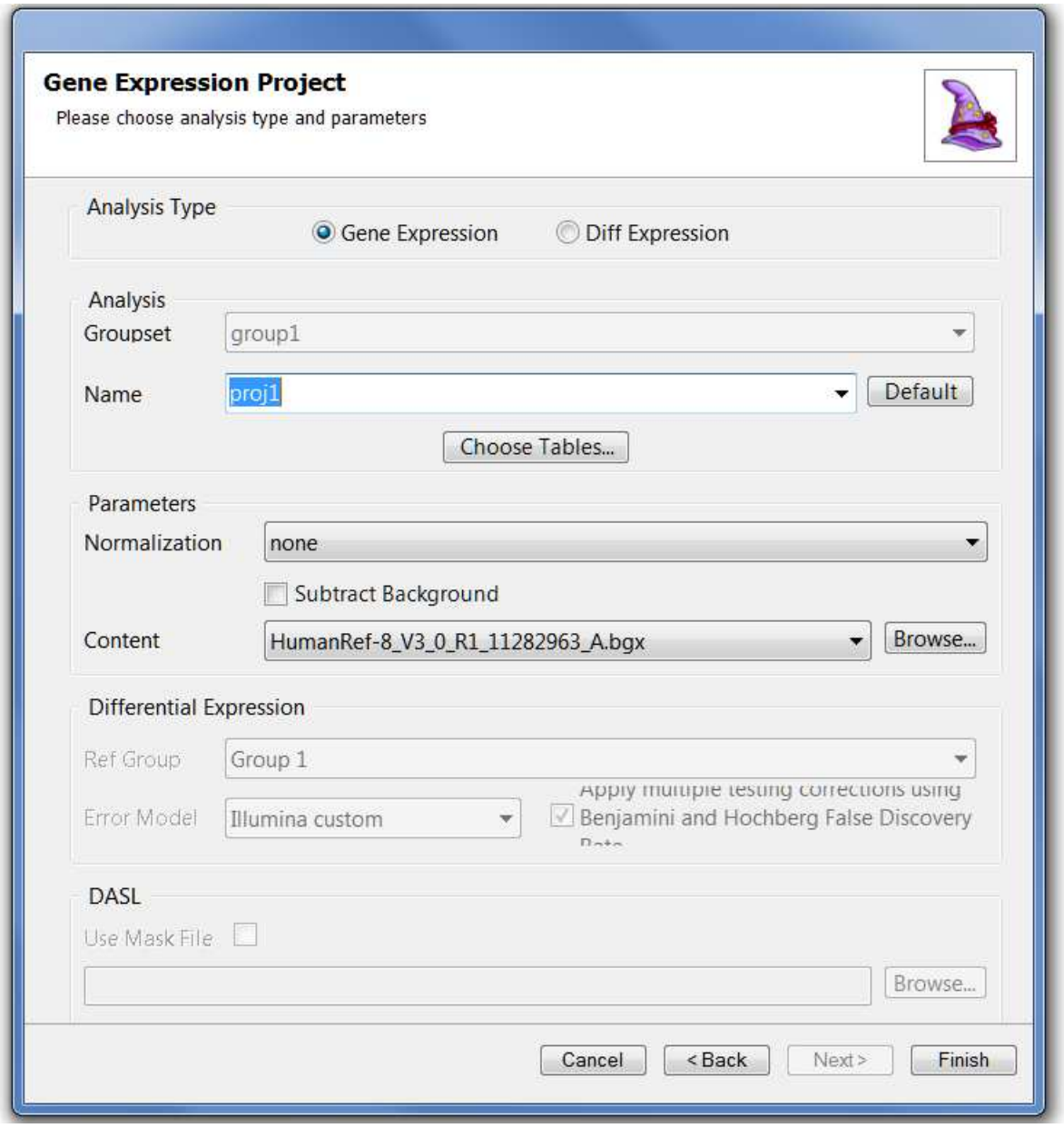
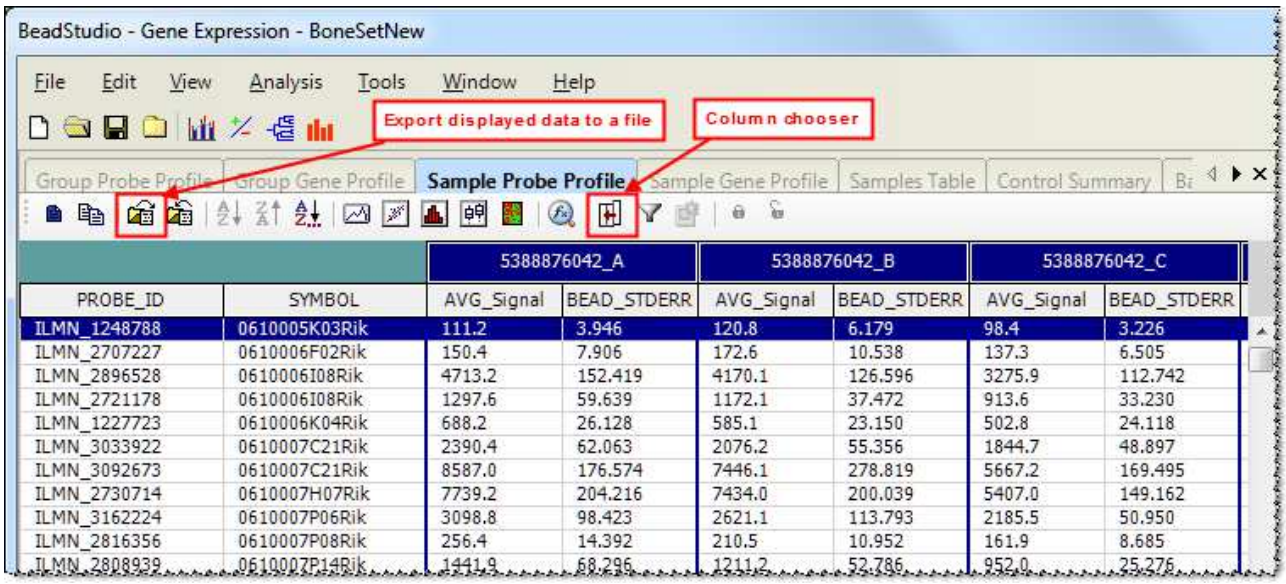


Fig.App.1 Configure Gene Expression Project in BeadStudio



		5388876042_A		5388876042_B		5388876042_C	
PROBE_ID	SYMBOL	AVG_Signal	BEAD_STDERR	AVG_Signal	BEAD_STDERR	AVG_Signal	BEAD_STDERR
ILMN_1248788	0610005K03Rik	111.2	3.946	120.8	6.179	98.4	3.226
ILMN_2707227	0610006F02Rik	150.4	7.906	172.6	10.538	137.3	6.505
ILMN_2896528	0610006I08Rik	4713.2	152.419	4170.1	126.596	3275.9	112.742
ILMN_2721178	0610006I08Rik	1297.6	59.639	1172.1	37.472	913.6	33.230
ILMN_1227723	0610006K04Rik	688.2	26.128	585.1	23.150	502.8	24.118
ILMN_3033922	0610007C21Rik	2390.4	62.063	2076.2	55.356	1844.7	48.897
ILMN_3092673	0610007C21Rik	8587.0	176.574	7446.1	278.819	5667.2	169.495
ILMN_2730714	0610007H07Rik	7739.2	204.216	7434.0	200.039	5407.0	149.162
ILMN_3162224	0610007P06Rik	3098.8	98.423	2621.1	113.793	2185.5	50.950
ILMN_2816356	0610007P08Rik	256.4	14.392	210.5	10.952	161.9	8.685
ILMN_2808939	0610007P14Rik	1441.9	68.296	1211.2	52.786	952.0	25.276

Fig.App.2 BeadStudio Gene Expression Analysis window – Sample Probe Profile tab

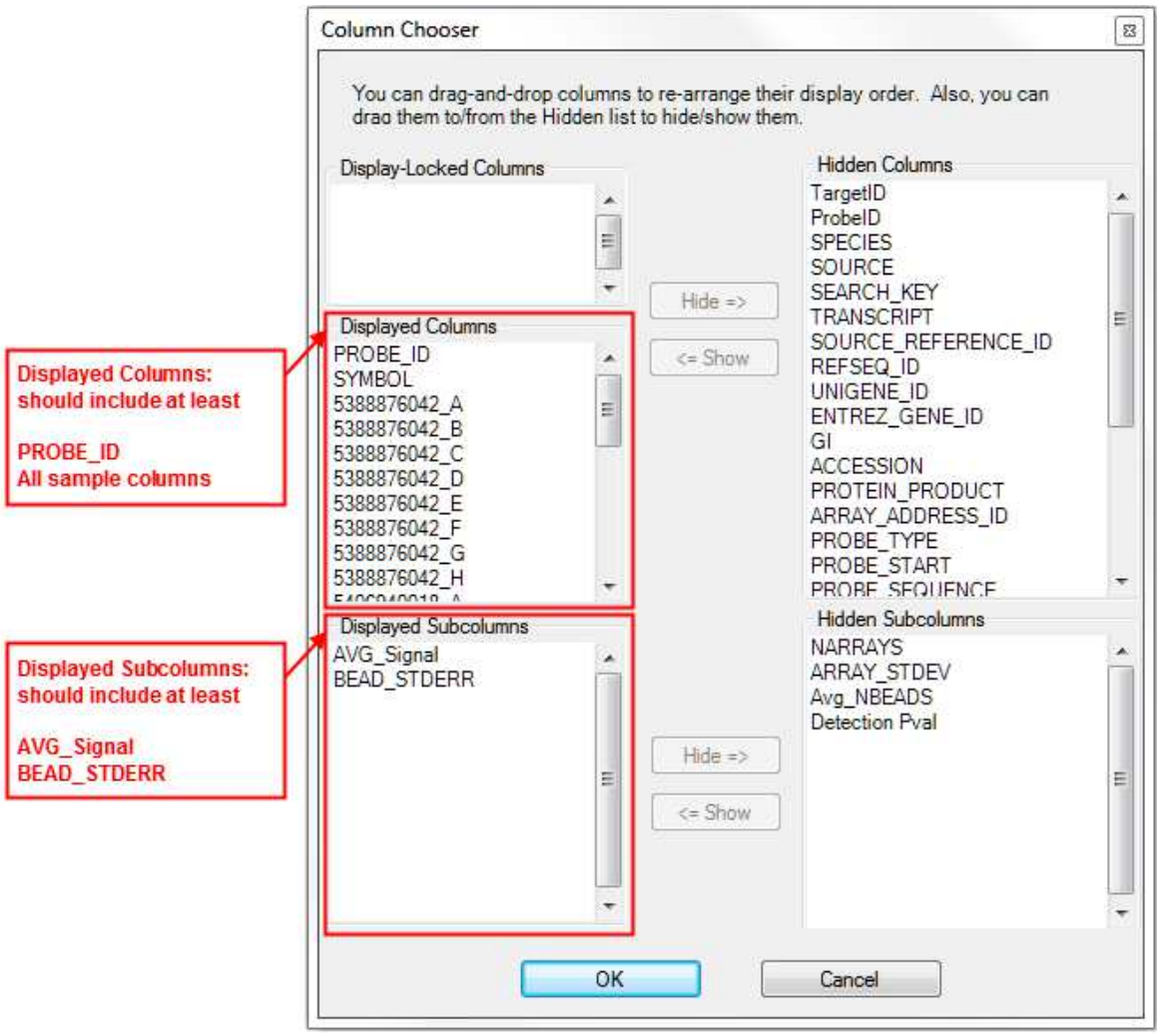
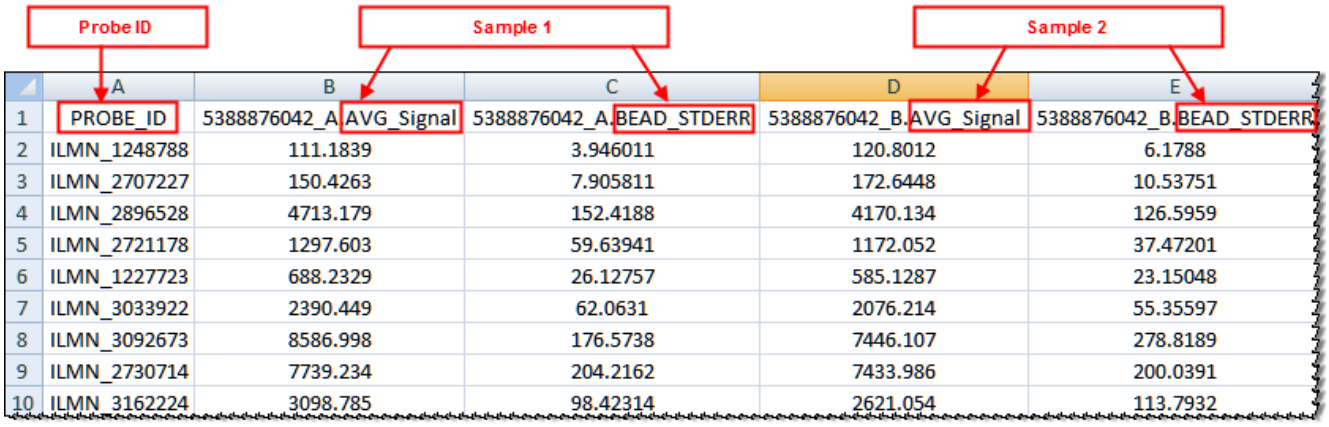
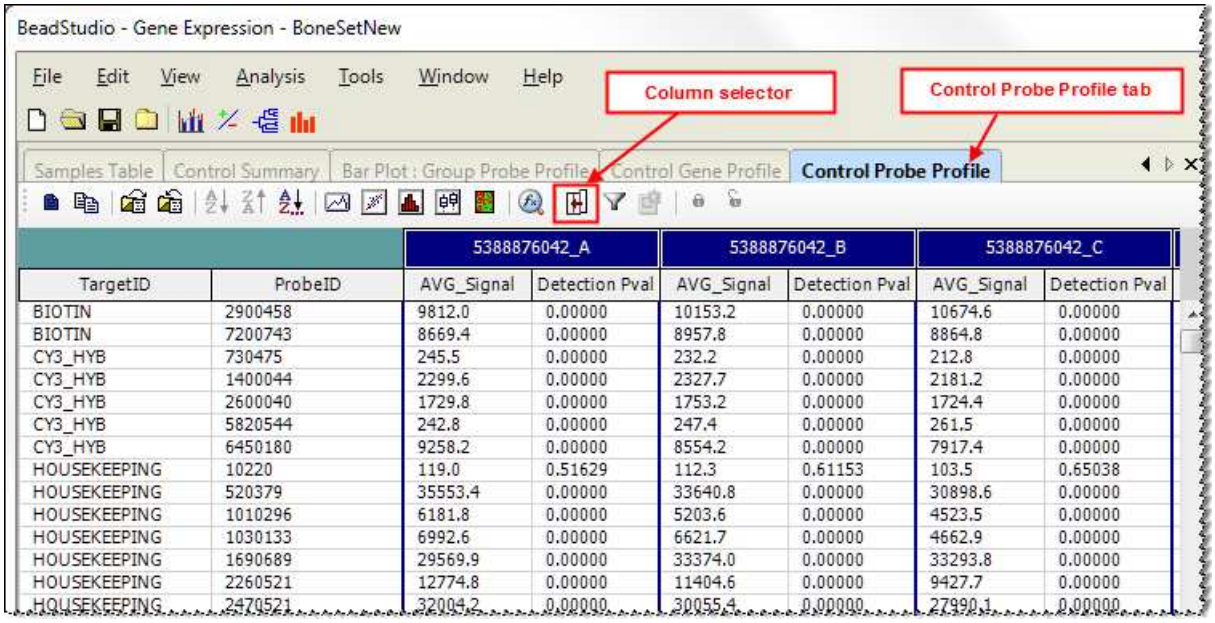


Fig.App.3 BeadStudio Column Chooser Window - Sample Probe Profile



	A	B	C	D	E
1	PROBE_ID	5388876042_A AVG_Signal	5388876042_A BEAD_STDERR	5388876042_B AVG_Signal	5388876042_B BEAD_STDERR
2	ILMN_1248788	111.1839	3.946011	120.8012	6.1788
3	ILMN_2707227	150.4263	7.905811	172.6448	10.53751
4	ILMN_2896528	4713.179	152.4188	4170.134	126.5959
5	ILMN_2721178	1297.603	59.63941	1172.052	37.47201
6	ILMN_1227723	688.2329	26.12757	585.1287	23.15048
7	ILMN_3033922	2390.449	62.0631	2076.214	55.35597
8	ILMN_3092673	8586.998	176.5738	7446.107	278.8189
9	ILMN_2730714	7739.234	204.2162	7433.986	200.0391
10	ILMN_3162224	3098.785	98.42314	2621.054	113.7932

Fig.App.4 Example of BeadStudio export data for Microarray Я US



		5388876042_A		5388876042_B		5388876042_C	
TargetID	ProbeID	AVG_Signal	Detection Pval	AVG_Signal	Detection Pval	AVG_Signal	Detection Pval
BIOTIN	2900458	9812.0	0.00000	10153.2	0.00000	10674.6	0.00000
BIOTIN	7200743	8669.4	0.00000	8957.8	0.00000	8864.8	0.00000
CY3_HYB	730475	245.5	0.00000	232.2	0.00000	212.8	0.00000
CY3_HYB	1400044	2299.6	0.00000	2327.7	0.00000	2181.2	0.00000
CY3_HYB	2600040	1729.8	0.00000	1753.2	0.00000	1724.4	0.00000
CY3_HYB	5820544	242.8	0.00000	247.4	0.00000	261.5	0.00000
CY3_HYB	6450180	9258.2	0.00000	8554.2	0.00000	7917.4	0.00000
HOUSEKEEPING	10220	119.0	0.51629	112.3	0.61153	103.5	0.65038
HOUSEKEEPING	520379	35553.4	0.00000	33640.8	0.00000	30898.6	0.00000
HOUSEKEEPING	1010296	6181.8	0.00000	5203.6	0.00000	4523.5	0.00000
HOUSEKEEPING	1030133	6992.6	0.00000	6621.7	0.00000	4662.9	0.00000
HOUSEKEEPING	1690689	29569.9	0.00000	33374.0	0.00000	33293.8	0.00000
HOUSEKEEPING	2260521	12774.8	0.00000	11404.6	0.00000	9427.7	0.00000
HOUSEKEEPING	2470521	32004.2	0.00000	30055.4	0.00000	27890.1	0.00000

FigApp.5 BeadStudio Gene Expression Analysis window - Control Probe Profile

I did not see the "Control Probe Profile" tab in the Gene Expression Analysis window.
The file may be invisible at default display settings. Click on "Window" in the menu bar and check all available files.

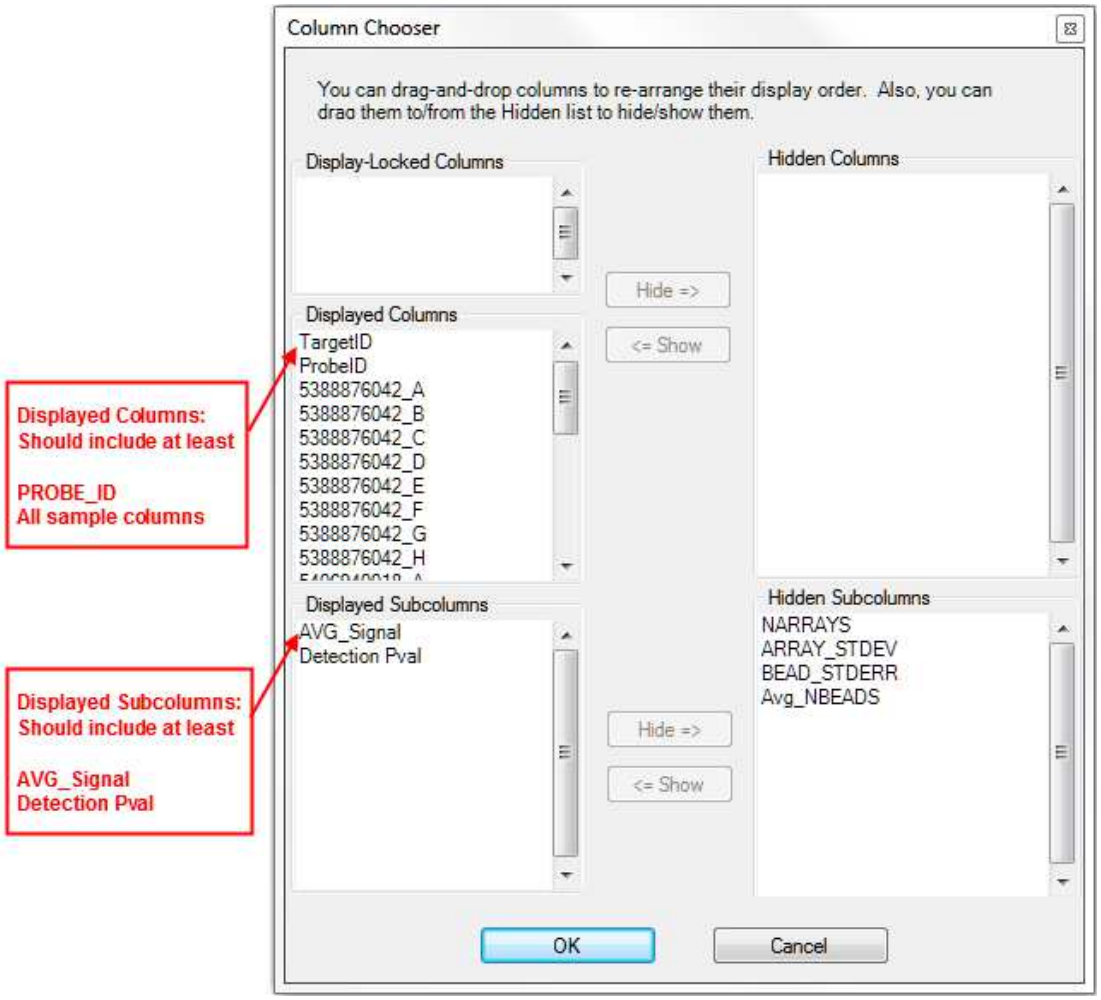


Fig.App.6 BeadStudio Column Chooser window - Control Probe Profile

> *Make sure TargetID is displayed before ProbeID!*
> *Otherwise, R will issue an error that prevents background correction procedure.*

APPENDIX 7. NOTES ON FOLDERS AND FILES

FOLDER MANAGEMENT IN MICROARRAY Я US

To help users to manage and track analysis results, we have made conscious efforts to automatically include key information into results file names, as well as create separate and clearly defined folder names to store different analysis results. The following diagram demonstrates folder management in Microarray Я US. Folder names are indicated by bold font. The procedure creating the corresponding folder is listed in parenthesis directly below.

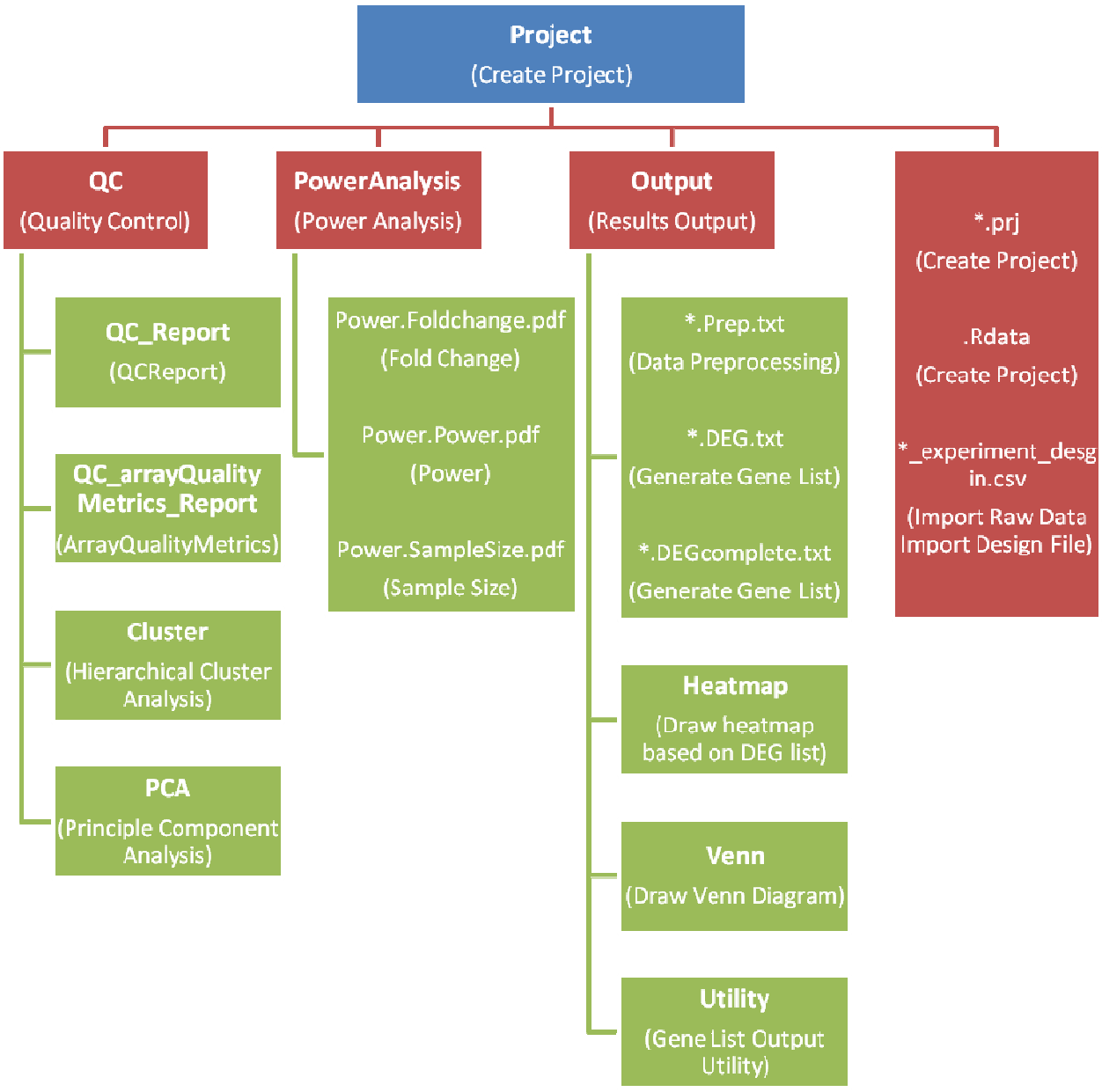


Fig.App.7 Folder management in Microarray Я US

FILE FORMATS

DESIGN FILE

A design file is a Comma Separated Value File (.csv) that specifies the experimental design information of your project.

- The first column MUST be named as "FileName" and includes all raw data file names (e.g. CEL file names for Affymetrix data or sample names for Illumina data) in the current project.
- The following columns should contain all major experimental attributes (one in each column). See example design file in Fig.App.8.
- Major experimental attribute includes all **consequential** factors, such as treatment, patient ID (for paired samples), time points (time course experiments), chip ID (for modeling batch effects), etc. Do NOT include irrelevant factors (e.g. chip platform, scanner).

	A	B	C	D
1	FileName	Genotype	Gender	
2	A.CEL	WT	F	
3	B.CEL	WT	F	
4	C.CEL	WT	M	
5	D.CEL	WT	M	
6	E.CEL	KO	F	
7	F.CEL	KO	F	
8	G.CEL	KO	M	
9	H.CEL	KO	M	
10				

First column must contain all file names in the current project

Fig.App.8 Example design file, design.csv

COMPLETE GENE LIST FILE (*.DEGcomplete.txt)

Complete gene list file includes differential expression analysis results for all probes on the chip (Fig.App.9). It is automatically generated by the **Generate Gene List** function in Microarray Я US. It can also be generated using external differential expression analysis results. To prepare the file in Excel, make sure that:

- The **first column** head must be named as **PROBEID** and contains all probe IDs on the chip. The order of the rest column will not affect any analysis.
- A **SYMBOL** column containing all official gene symbols must be included.
- Although not required, it is highly recommended to name **p-value column as P** and a **fold-change column as FC**. Missing such information may result in generating incorrectly

formatted files for certain functional enrichment program when using the **Gene List Output Utility**.

- Users may include any additional information in other columns.
- Save the file as a tab delimited file, and name it as *.DEGcomplete.txt

DEG LIST FILE (*.DEG.txt)

DEG list file includes analysis results for **differentially expressed probes** (i.e. probes that passed user specified cutoffs). It is a sub-file of its corresponding Complete gene list file and follows the same format (Fig.App.9). It can be automatically generated by the **Generate Gene List** function in Microarray Я US. It can be prepared using external differential expression analysis results following the same instructions for generating Complete gene list files.

The first column must be named as PROBEID and contain probe IDs
Other required columns are marked as red and must be named as indicated

PROBEID	SYMBOL	ENTREZID	Case.Mean	Control.Mean	P	FC
1415670_at	Copg	54161	9.475	9.470	0.927074823	1.00
1415671_at	Atp6v0d1	11972	12.304	12.076	0.036499877	1.17
1415672_at	Golga7	57437	11.728	11.922	0.005109026	-1.14
1415673_at	Psph	100678	8.980	9.055	0.23032531	-1.05
1415674_a_at	Trappc4	60409	10.471	11.081	5.3793E-10	-1.53
1415675_at	Dpm2	13481	10.225	10.317	0.058302317	-1.07
1415676_a_at	Psmb5	19173	11.761	11.791	0.411769005	-1.02
1415677_at	Dhrs1	52585	9.880	9.719	0.05452822	1.12
1415678_at	Ppm1a	19042	11.279	11.503	0.001468586	-1.17

Fig.App.9 Example: Complete gene list and DEG list file

PREPROCESSED EXPRESSION FILE (*.Prep.txt)

Preprocessed expression file includes preprocessing results for all probes on a chip (Fig.App.10). It is automatically generated by the **Data Preprocessing** function in Microarray Я US. It can also be prepared using external preprocessing analysis results. To prepare the file in Excel, make sure that:

- The **first column** must be named as **PROBEID** and contains all probe IDs on the chip.
- Each of the following columns contains preprocessed expression intensities for each sample. Using the same sample names (or CEL file names) as listed in the FileName column in the Design file (Fig.App.8).
- Save the file as tab delimited file, and name it as *.Prep.txt

The first column must be named as PROBEID and contain all probes IDs on the array
The rest of the column must be named as sample names and contain preprocessed expression intensities

PROBEID	A.CEL	B.CEL	C.CEL	D.CEL	E.CEL	F.CEL	G.CEL	H.CEL
1415670_at	9.561104216	9.06651524	9.114952809	9.196444858	9.23057228	9.506547131	9.63759429	9.568194544
1415671_at	11.91924088	11.71950539	11.63765774	11.95723136	11.86955277	12.5177369	12.54369887	12.44319279
1415672_at	11.65911755	11.41971763	11.95602589	12.06401375	12.00584291	11.86231971	11.82782046	11.81163665
1415673_at	8.743939736	8.631228819	9.060426701	9.082292577	9.066490765	9.103090379	8.878674841	9.003904172
1415674_a_at	10.79914341	10.76710854	11.20028277	11.19898709	11.21942966	10.37985738	10.47654262	10.42919101
1415675_at	10.21522001	10.10249871	10.34256811	10.28779386	10.20907777	10.20288246	10.28535374	10.24578954
1415676_a_at	11.66201748	11.21250296	11.55077616	11.79071963	11.78798676	11.77282867	11.79704315	11.73138139
1415677_at	9.654382098	9.464064873	9.581922771	9.56096836	9.562028656	10.00317222	9.953072002	9.878125224
1415678_at	11.29655886	11.18796132	11.54299759	11.54152498	11.54212385	11.33090966	11.24597191	11.21300078

Fig.App.10 Example: Preprocessed expression file

APPENDIX 8. TUTORIAL FOR PREPARING PARTEK GENOMICS SUITE (PARTEK GS) ANALYSIS RESULTS TO USE THE GENE LIST OUTPUT UTILITY

To use the gene list output utility with Partek GS (Partek Inc. St. Louis, MO) analysis results, prepare the required files in the Partek GS. Refer to the previous "File Format" session for more details.

THE PREPROCESSED EXPRESSION FILE:

- Select the imported data (preprocessed data) datasheet in Partek, From File>>Transform, select "Create Transposed Spreadsheet" to create a new spreadsheet.
- Save this new spreadsheet as *.DEGcomplete.txt.

THE COMPLETE GENE LIST FILE:

- Select the ANOVAresult datasheet in Partek and make sure that the Gene_Symbol Columns is included. If missing, use the Insert Annotation function (right mouse click while selecting the Probeset ID column>>Insert Annotation) to add it in.
- Save this spreadsheet as *.DEGcomplete.txt.
- Open the tab delimited file in Excel
- Move "Probe ID" column to the first column and rename it as "PROBEID". Rename the "Gene_Symbol", "p-value (exp vs. ctrl)"and "fold-change (exp vs. ctrl)" columns to be "SYMBOL", "P" and "FC", respectively.
- Save the changes

THE DEG LIST FILE:

- Select the desired gene list spreadsheet in Partek and make sure that the Gene_Symbol Columns is included. If missing, use the Insert Annotation function to add it in.
- Save this spreadsheet as *.DEG.txt.
- Open the tab delimited file in Excel
- Move "Probe ID" column to the first column and rename it as "PROBEID". Rename the "Gene_Symbol", "p-value (exp vs. ctrl)"and "fold-change (exp vs. ctrl)" columns to be "SYMBOL", "P" and "FC", respectively.
- Save the changes

THE DESIGN FILE:

- Prepare the design file manually using Excel. Refer to File Format section in **Appendix 7: Notes on Folders and Files**.

REFERENCES

Affymetrix (2002). "GeneChip Expression Analysis: Data Analysis Fundamentals."

Al-Shahrour, F., P. Minguez, et al. (2007). "FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments." Nucl. Acids Res. **35**(suppl_2): W91-96.

Antonov, A. V., S. Dietmann, et al. (2009). "GeneSet2miRNA: finding the signature of cooperative miRNA activities in the gene lists." Nucl. Acids Res. **37**(suppl_2): W323-328.

Backes, C., A. Keller, et al. (2007). "GeneTrail--advanced gene set enrichment analysis." Nucl. Acids Res. **35**(suppl_2): W186-192.

Barbosa-Morais, N. L., M. J. Dunning, et al. (2010). "A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data." Nucleic Acids Research **38**(3): e17.

Barrett, T., D. B. Troup, et al. (2011). "NCBI GEO: archive for functional genomics data sets--10 years on." Nucleic Acids Research **39**(Database issue): D1005-1010.

Berriz, G. F., J. E. Beaver, et al. (2009). "Next generation software for functional trend analysis." Bioinformatics **25**(22): 3043-3044.

Breitling, R., P. Armengaud, et al. (2004). "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments." FEBS letters **573**(1-3): 83-92.

Conesa, A., M. J. Nueda, et al. (2006). "maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments." Bioinformatics **22**(9): 1096-1102.

Dai, M., P. Wang, et al. (2005). "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." Nucleic Acids Research **33**(20): e175.

Development Core Team (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing.

Draghici, S., P. Khatri, et al. (2007). "A systems biology approach for pathway level analysis." Genome Res. **17**(10): 1537-1545.

Du, P., W. A. Kibbe, et al. (2007). "nuID: a universal naming scheme of oligonucleotides for illumina, affymetrix, and other microarrays." Biol Direct **2**: 16.

Du, P., W. A. Kibbe, et al. (2008). "lumi: a pipeline for processing Illumina microarray." Bioinformatics **24**(13): 1547-1548.

Eden, E., R. Navon, et al. (2009). "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists." BMC Bioinformatics **10**(1): 48.

Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Research **30**(1): 207-210.

Essaghir, A., F. Toffalini, et al. (2010). "Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data." Nucleic Acids Research **38**(11): e120.

Gautier, L., L. Cope, et al. (2004). "affy--analysis of Affymetrix GeneChip data at the probe level." Bioinformatics **20**(3): 307-315.

Gentleman, R. C., V. J. Carey, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**(10): R80.

Hong, F., B. Wittner, et al. (2009). "RankProd: Rank Product methods for identifying differentially expressed genes with application in meta-analysis." R version 2.20.0.

Huang da, W., B. T. Sherman, et al. (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic Acids Research **37**(1): 1-13.

Huang da, W., B. T. Sherman, et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nature protocols **4**(1): 44-57.

Huber, W., A. von Heydebreck, et al. (2002). "Variance stabilization applied to microarray data calibration and to the quantification of differential expression." Bioinformatics **18 Suppl 1**: S96-104.

Irizarry, R. A., B. Hobbs, et al. (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics **4**(2): 249-264.

Kaimal, V., E. E. Bardes, et al. (2010). "ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems." Nucleic Acids Research **38**(Web Server issue): W96-102.

Kauffmann, A., R. Gentleman, et al. (2009). "arrayQualityMetrics--a bioconductor package for quality assessment of microarray data." Bioinformatics **25**(3): 415-416.

Kauffmann, A., T. F. Rayner, et al. (2009). "Importing ArrayExpress datasets into R/Bioconductor." Bioinformatics **25**(16): 2092-2094.

Kielbasa, S. M., H. Klein, et al. (2010). "TransFind--predicting transcriptional regulators for gene sets." Nucleic Acids Research **38**(Web Server issue): W275-280.

Kuehn, H., A. Liberzon, et al. (2008). "Using GenePattern for gene expression analysis." Curr Protoc Bioinformatics **Chapter 7**: Unit 7 12.

Lamb, J. (2007). "The Connectivity Map: a new tool for biomedical research." Nat Rev Cancer **7**(1): 54-60.

Li, C. and W. H. Wong (2001). "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection." Proceedings of the National Academy of Sciences of the United States of America **98**(1): 31-36.

Lin, S. M., P. Du, et al. (2008). "Model-based variance-stabilizing transformation for Illumina microarray data." Nucleic Acids Res **36**(2): e11.

Nam, S., M. Li, et al. (2009). "MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression." Nucl. Acids Res. **37**(suppl_2): W356-362.

Nogales-Cadenas, R., P. Carmona-Saez, et al. (2009). "GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information." Nucl. Acids Res. **37**(suppl_2): W317-322.

Parkinson, H., U. Sarkans, et al. (2011). "ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments." Nucleic Acids Research **39**(Database issue): D1002-1004.

Parman, C., C. Halling, et al. (2010). "affyQCReport: QC Report Generation for affyBatch objects." R package version 1.26.0.

Prieto, C., A. Risueno, et al. (2008). "Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles." PLoS One **3**(12): e3911.

Risueno, A., C. Fontanillo, et al. (2010). "GATExplorer: genomic and transcriptomic explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs." BMC Bioinformatics **11**: 221.

Sales, G., A. Coppe, et al. (2010). "MAGIA, a web-based tool for miRNA and Genes Integrated Analysis." Nucleic Acids Research **38**(Web Server issue): W352-359.

Salomonis, N., K. Hanspers, et al. (2007). "GenMAPP 2: new features and resources for pathway analysis." BMC Bioinformatics **8**(1): 217.

Schwender, H. (2009). "siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches." R package version 1.22.0.

Sean, D. and P. S. Meltzer (2007). "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." Bioinformatics **23**(14): 1846-1847.

Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Stat Appl Genet Mol Biol **3**: Article3.

Subramanian, A., H. Kuehn, et al. (2007). "GSEA-P: a desktop application for Gene Set Enrichment Analysis." Bioinformatics **23**(23): 3251-3253.

Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proceedings of the National Academy of Sciences of the United States of America **98**(9): 5116-5121.

Warnes, G. R., P. Liu, et al. (2009). "ssize: Estimate Microarray Sample Size." R package version 1.22.0.

Wu, J., Q. Qiu, et al. (2009). "Web-based interrogation of gene expression signatures using EXALT." BMC Bioinformatics **10**(1): 420.

Wu, J. Z., R. Irizarry, et al. (2002). "gcrma: Background Adjustment Using Sequence Information." R package version 2.20.0.

Wu, Z., R. Irizarry, et al. (2004). "A Model Based Background Ajustment for Oligonucleotide Expression Arrays." Journal of American Statistical Association **99**(468): 909-917.

Wu, Z. and R. A. Irizarry (2005). "Stochastic models inspired by hybridization theory for short oligonucleotide arrays." Journal of computational biology : a journal of computational molecular cell biology **12**(6): 882-893.

Zeeberg, B. R., W. Feng, et al. (2003). "GoMiner: a resource for biological interpretation of genomic and proteomic data." Genome Biology **4**(4): R28.

Zhang, B., S. Kirov, et al. (2005). "WebGestalt: an integrated system for exploring gene sets in various biological contexts." Nucleic Acids Research **33**(suppl_2): W741-748.

Zhu, Y., S. Davis, et al. (2008). "GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus." Bioinformatics **24**(23): 2798-2800.