# Microarray Data Analysis Using BRB-ArrayTools Version 4.3.0 Beta3

Eric Polley, Lori Long, Supriya Menezes

Mar 4th, 5th 2013

# Agenda
# Day 1

I. What is BRB-ArrayTools?
II. Installing BRB-ArrayTools and its required components
III. Creating a collated project workbook
IV. Data filtering and normalization options
V. Break
VI. Graphics
VII. Class Comparison
VIII. Gene Set comparison

# Agenda
# Day 2

I. Clustering
II. MDS
III. Class Prediction
IV. Plug-ins
V. Tutorial.
VI. Hands-on.

# Part I:

# What is BRB-ArrayTools?

# BRB-ArrayTools

## An Integrated Software Tool for Microarray Data Analysis

- Developed under the direction of Dr. Richard Simon of the Biometrics Research Branch, NCI.
- Software was developed with the purpose of deploying powerful statistical tools for use by biologists.
- Analyses are launched from user-friendly Excel interface. Also requires installation of a free software called R for running back-end programs. Current requirement for R is v 2.15.2. Publicly available from BRB website:

http://linus.nci.nih.gov/BRB-ArrayTools.html

# Features of BRB-ArrayTools

- Capability to collate (sort into an expression data matrix) microarray data from a set of experiments, and apply filtering and normalization.
- The focus of the software has been the implementation of statistical methodology which utilizes the sample descriptors (supervised analysis).
- Scatterplots, hierarchical clustering, and multidimensional scaling analyses also provide powerful visualization tools.
- Gene annotations are integrated into analysis output to inform the analysis results. Also, includes analyses using BioCarta,KEGG and Broad/MIT pathways.
- Advanced users may program their own plug-in analysis tools within BRB-ArrayTools.

# Limitations of BRB-ArrayTools

- Available only on the PC. As well as on an Apple macbook pro machine with Windows OS installed with Apple's bootcamp software/Parallels.
- Currently compatible with MS Vista/ Windows 7 and Excel 2007/2010.
- Importing and running analysis tools on large data sets may require a large memory capacity and be time-consuming.

# Recently Added new tools

- Importing Illumina methylation data.
- Importing RNA-Seq data pre-processed using the Galaxy web tools (https://main.g2.bx.psu.edu/)
- Heatmap of data to generate a heatmap on clustered data to provide users an overview on their data.
- Zoomable SVG format heatmap.
- Two plug-ins specifically for methylation data.

# Installing BRB-ArrayTools

- http://linus.nci.nih.gov/BRB-ArrayTools.html
- Register to obtain a user name and password by going to the guestbook.
- Select the version you wish to download.
- Currently available BRB-ArrayToolsv4.2.1 and v4.3.0 beta2.

# BRB-ArrayTools Web Page

## BRB-ArrayTools

### Developed by: Richard Simon & BRB-ArrayTools Development Team

BRB-ArrayTools is an integrated package for the visualization and statistical analysis of DNA microarray gene expression data. It was developed by professional statisticians experienced in the analysis of microarray data and involved in the development of improved methods for the design and analysis of microarray based experiments. The array tools package utilizes an Excel front end. Scientists are familiar with Excel and utilizing Excel as the front end makes the system portable and not tied to any database. The input data is assumed to be in the form of Excel spreadsheets describing the expression values and a spreadsheet providing user-specified phenotypes for the samples arrayed. The analytic and visualization tools are integrated into Excel as an add-in. The analytic and visualization tools themselves are developed in the powerful R statistical system, in C and Fortran programs and in Java applications. Visual Basic for Applications is the glue that integrates the components and hides the complexity of the analytic methods from the user. The system incorporates a variety of powerful analytic and visualization tools developed specifically for microarray data analysis.


Download version 4.2.1 Stable Release
(Released on January 24, 2012 )

All required components in ONE file
Individual components


FAQs & Answers

BRB-ArrayTools Message Board

BRB-ArrayTools Data Archive
for Human Cancer Gene Expression

Email BRB-ArrayTools Support


Download version 4.3.0 Beta2 Release
(Released on Nov 28, 2012 ) What's New

All required components in ONE file
Individual components

Book for DNA Microarray Analysis

Publications Based on BRB-ArrayTools
Analyses

BRB-ArrayTools User Community
Institution List

# Full Installer

- Also available is an option to download a FULL installer. This file is a bundle of all the necessary components like Rv2.15.2 and java are included along with ArrayTools and CGHTools.

# Installing BRB-ArrayTools Pre-download


Biometric Research Branch

*Division of Cancer Treatment and Diagnosis*

**Before installing BRB-ArrayTools, please download, and install the following two software packages IN THE ORDER GIVEN BELOW. If you already have them installed, please click here to go to BRB-Ar**

1. Download and install Java from http://www.java.com

2. Download and install R 2.15.2 from http://cran.r-project.org/

Go to BRB-ArrayTools download page

Guest Book | Message Board | Download | Feedback | Licenses | Reprints and Presentations

# Downloading BRB-ArrayTools

- After installing the necessary components like R, statconnDCOM and Java, download and install BRB-ArrayTools.

BRB-ArrayTools Download Page

Developed by: Richard Simon & BRB-ArrayTools Development Team

*If you are a new user, please complete registration at our Guest Book before installation. This software is free for non-commercial use. Commercial users should contact Micho licensing. For technical issues, please contact BRB-ArrayTools Support at arraytools@emmes.com.

Download version 4.3.0 beta2 release

*Instructions for Excel 2007 and 2010 Users to set security level and load the Add-Ins into Excel 2007 and 2010 after installation.

The following documentation files are included in the above software installations, or may be downloaded separately for perusal prior to installation of the software.

Download Readme file

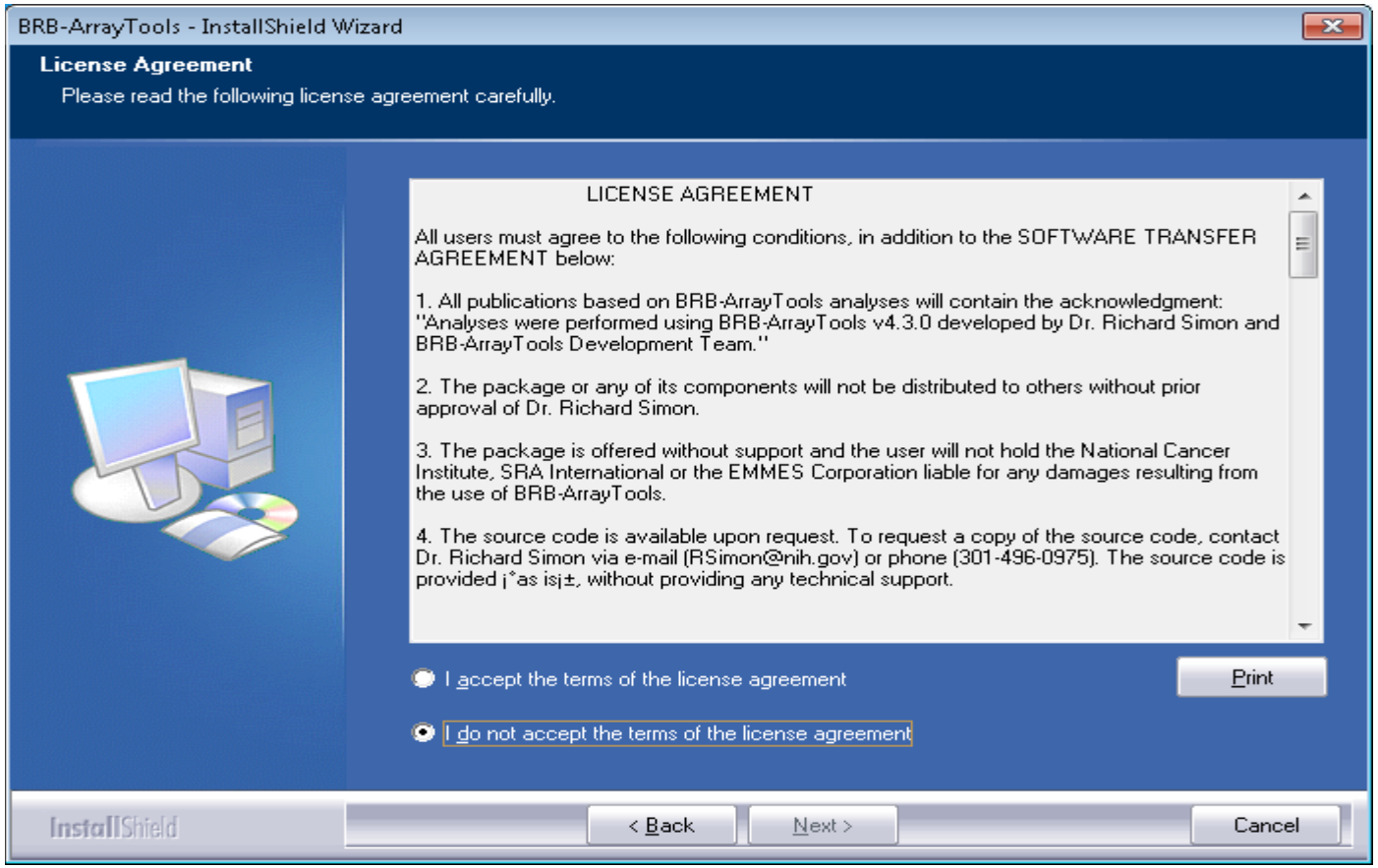
Download BRB ArrayTools Manual

Download the new Scatter Plot demo video

# Installing BRB-ArrayTools

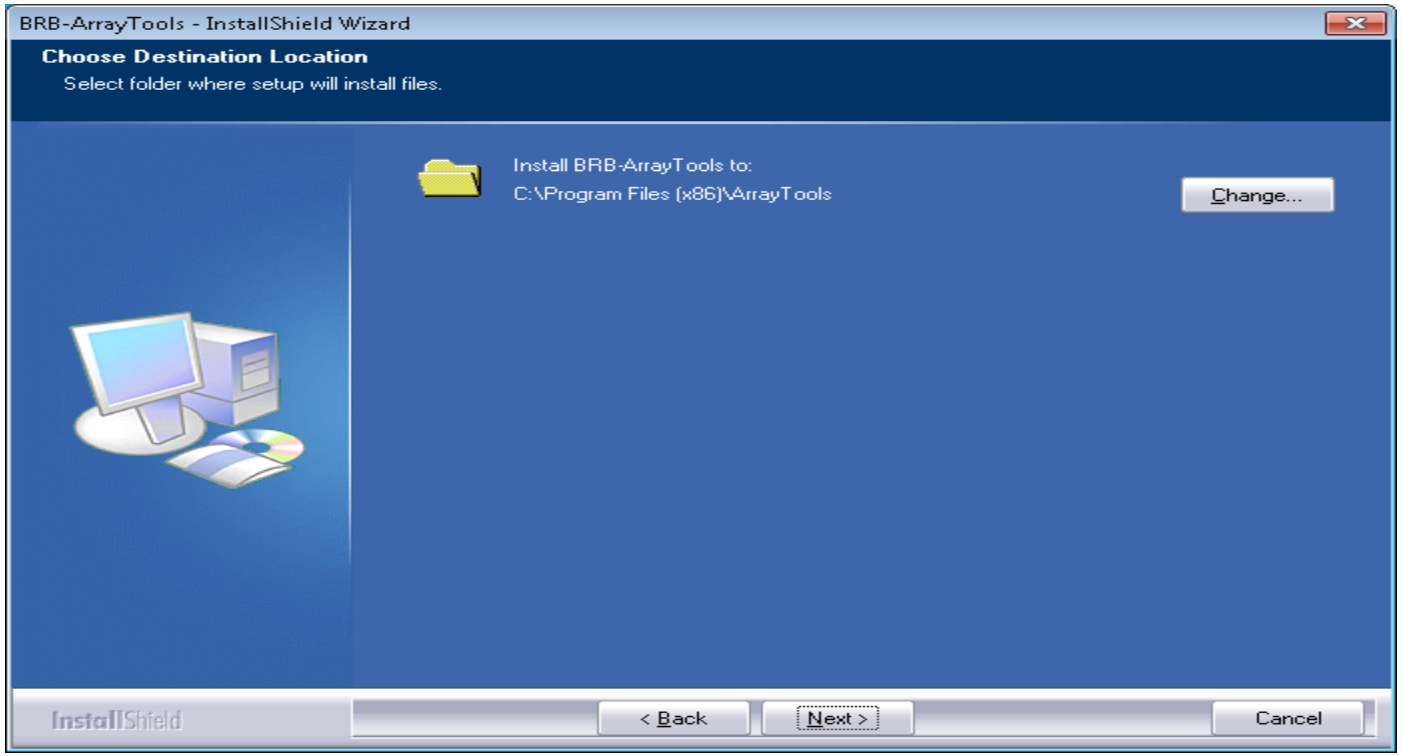
- On your desktop look for the folder called “BRB-ArrayTools-Class”.
- The ALL in ONE file called “ArrayTools_v4_3_0_Beta_3_Full.exe”.

# Installing BRB-ArrayTools

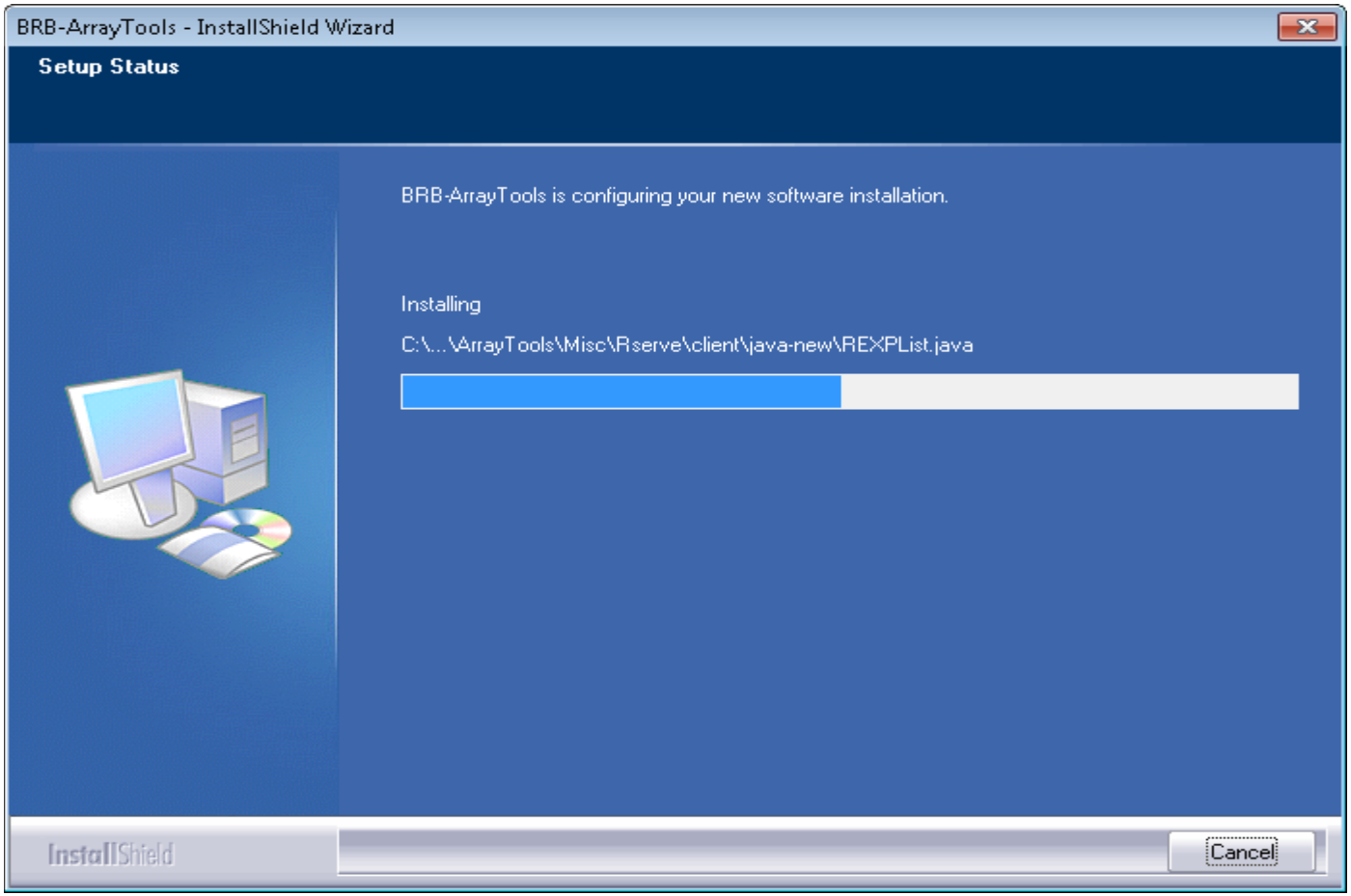


# Installing BRB-ArrayTools

- Select installation folder.

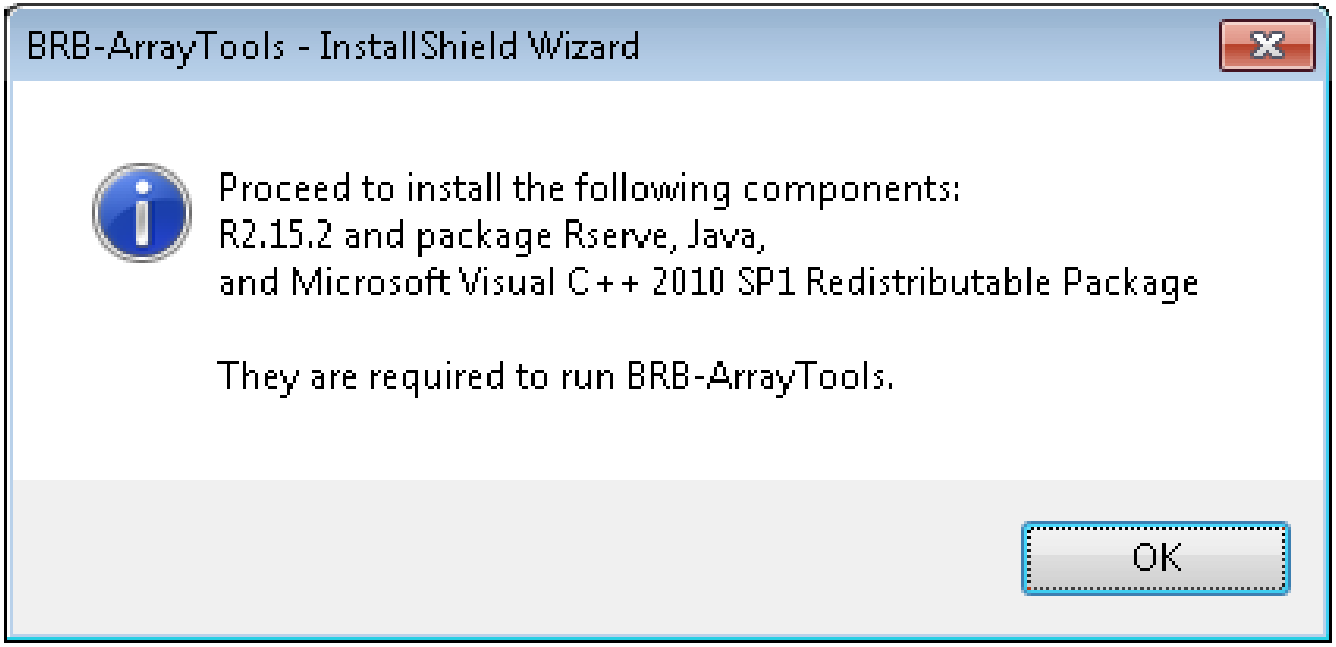


# Installing BRB-ArrayTools



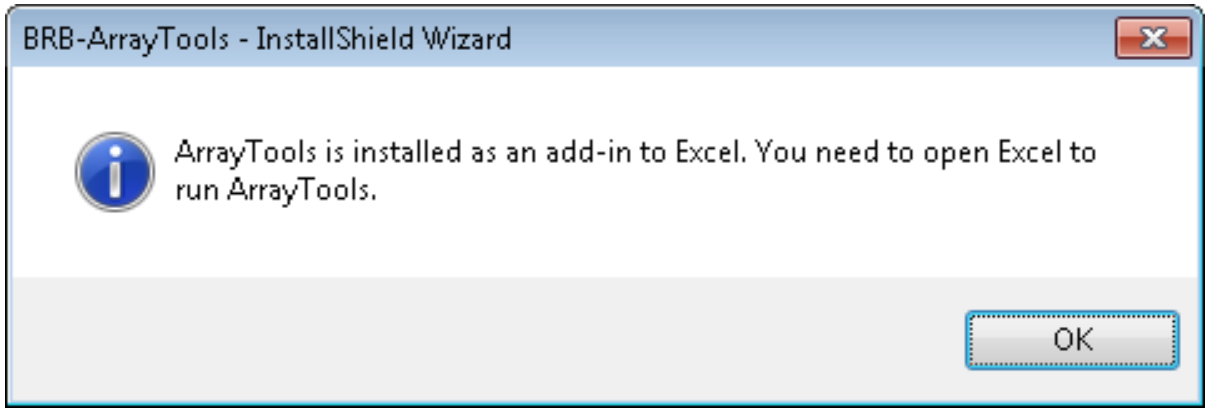
# Installing BRB-ArrayTools

- Proceed to install required components



# Installing BRB-ArrayTools

- After successfully installing BRB-ArrayTools, you will be prompted with the message below.
- Click “OK” as the software has been installed as an add-in to Excel.



# Excel 2007/2010- loading the add-in

- 1:Click on the Microsoft ‘Office’ button on the top left corner of the Excel menu.
- 2. Then, select the “Excel Options” button on the bottom right.
- 3: Click on “Trust Center”
- 4. Then click on “Trust Center Settings”
- 5: Choose the “Macro Settings” from the left hand panel.
- 6. Check “Enable all macros” and “Trust access to VBA project.”
- 7. Click the “OK” button.
- 8: Choose the “Add-ins” option from the left hand tab.
- 9. Click “BRB-ArrayTools” on the Active or Inactive application add-in.
- 10. Hit the “Go” button down at the bottom.
- 11. Check all the three “Add-ins”, BRB-ArrayTools and CGHTools.
- 12. Then click OK.
- If you don’t see the “Add-ins” ribbon along side “Home Insert…Review View” panel at the top then please close Excel and re-start.
- On clicking on Add-Ins tab, the Add-Ins should be listed there namely: ArrayTools and CGHTools add-ins.

# [Hands-on instructions]

## [Getting started]

1. Open **Excel**.
2. Click on **Tools →Add-ins**, and see that **BRB-ArrayTools** is loaded as an add-in.
3. When BRB-ArrayTools is loaded as an add-in, you will find an **ArrayTools** menu. This is the interface for all BRB-ArrayTools functions.
4. Click on **ArrayTools → Getting started**.
5. Here you will see the **Tutorial** and **Open a sample dataset** options.
6. For Office 2007, click on the “Add-ins” and you should find “ArrayTools”.

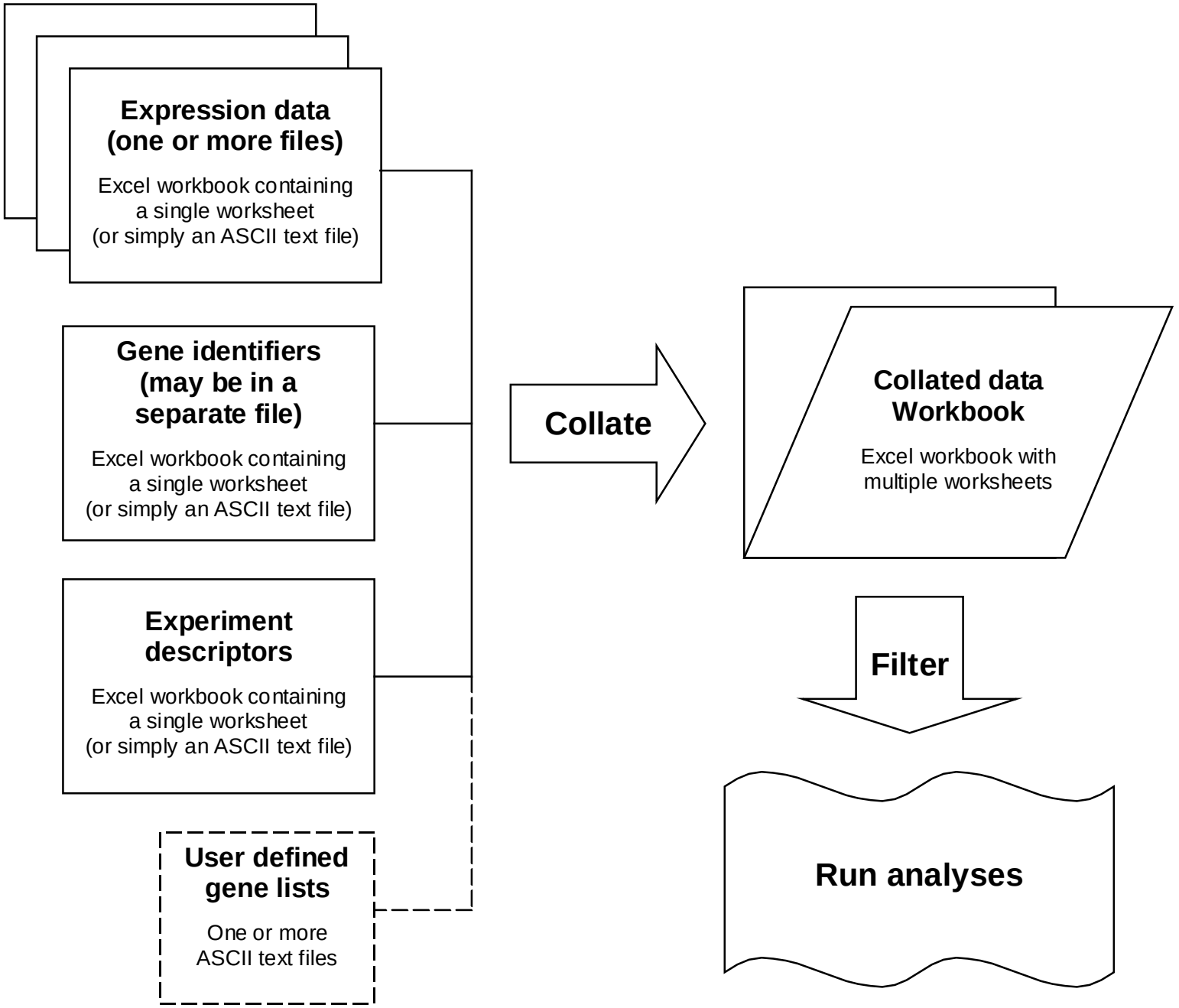
# ArrayTools Menu

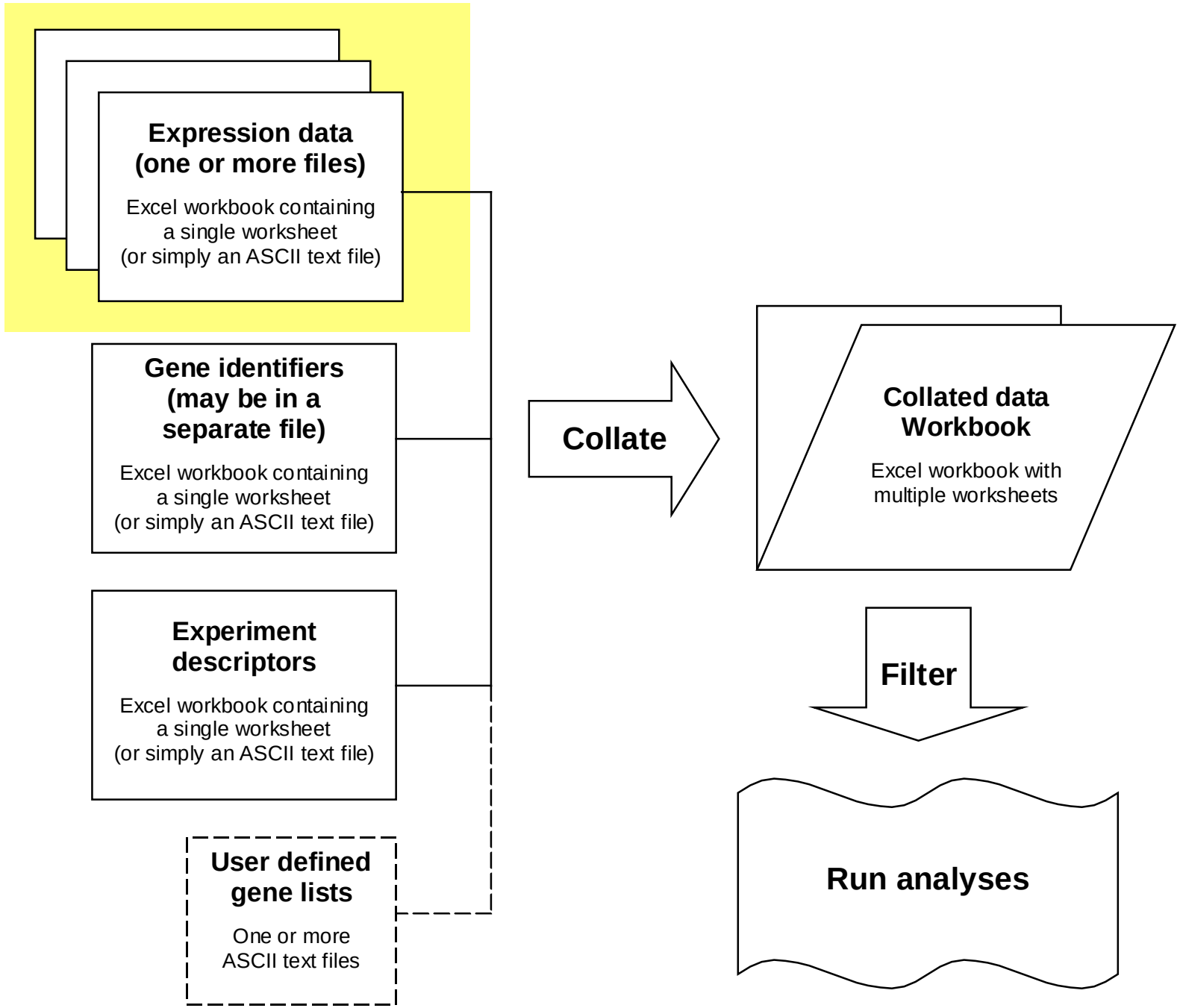
- From the “ArrayTools” pull down menu, you can navigate to the following:

- Manuals
- Tutorials
- Utilities
- Support
- License and version information

# Part II:

# Getting your data into BRB-ArrayTools: Creating a project workbook

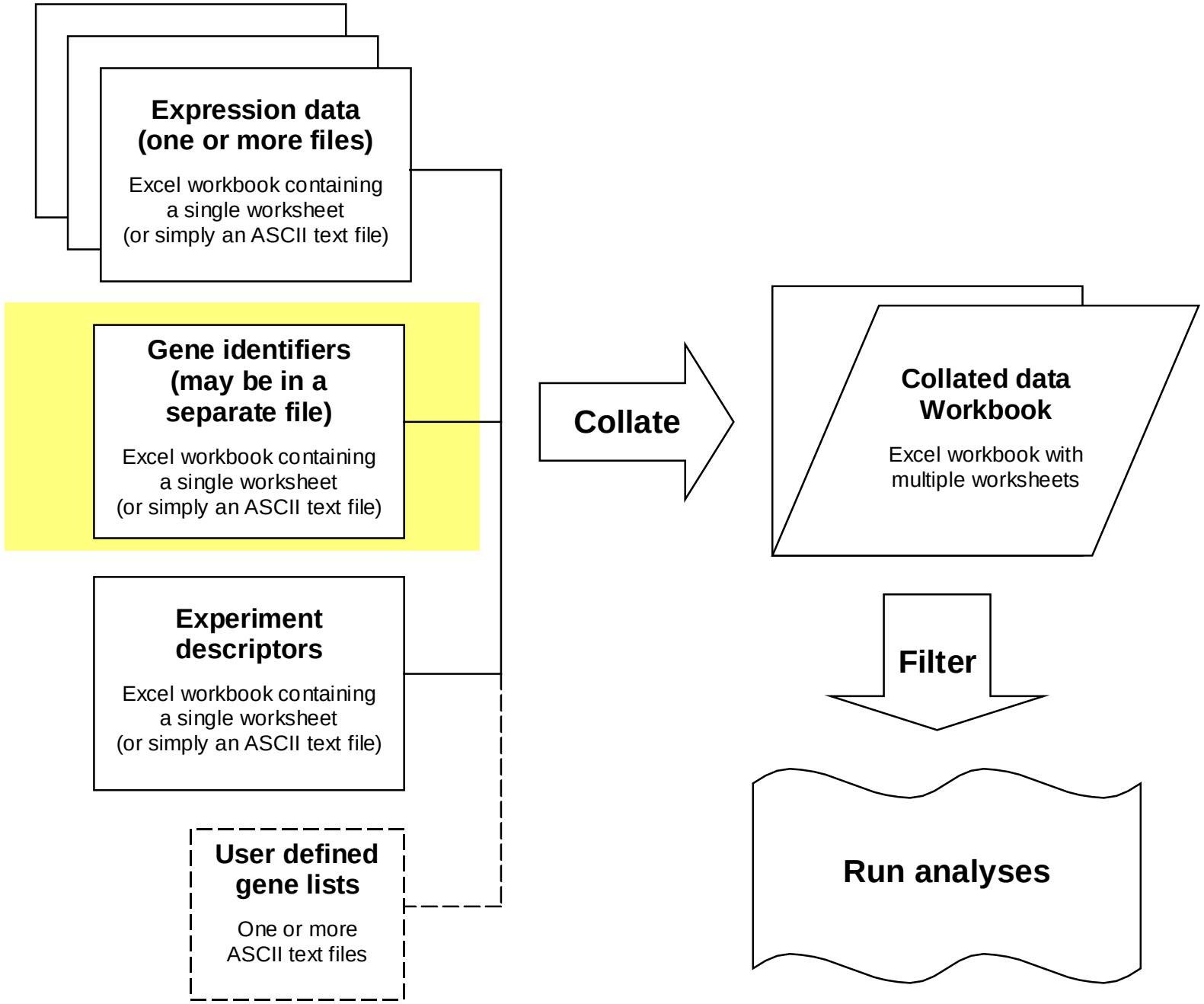

Expression data
(one or more files)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Gene identifiers
(may be in a
separate file)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Experiment
descriptors
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
User defined
gene lists
One or more
ASCII text files
Collate
Collated data
Workbook
Excel workbook with
multiple worksheets
Filter
Run analyses


Expression data
(one or more files)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Gene identifiers
(may be in a
separate file)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Experiment
descriptors
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
User defined
gene lists
One or more
ASCII text files
Collate
Collated data
Workbook
Excel workbook with
multiple worksheets
Filter
Run analyses

# Expression data

- Input data as tab-delimited ASCII files (or Excel spreadsheets) in one of the following formats:
    1. Horizontally aligned
    2. Separate files

- Files may contain expression data in the form of signal (or single-channel expression summary), dual-channel intensities, or expression ratios (for dual-channel data). Data may or may not have been already log-transformed. Flags, detection call, and spot size may also be used. All other variables will be ignored.


Expression data
(one or more files)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Gene identifiers
(may be in a
separate file)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Experiment
descriptors
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
User defined
gene lists
One or more
ASCII text files
Collate
Collated data
Workbook
Excel workbook with
multiple worksheets
Filter
Run analyses

# Gene identifiers

- A gene identifiers file is optional, but highly recommended for annotation purposes.
- Gene identifiers which may be used for hyperlinking are:  clone ids, UniGene cluster id or gene symbol, GenBank/RefSeq accessions, Entrez IDs and probe set ids.

# Gene identifiers

Two examples of a gene identifier file

GeneIds.xls

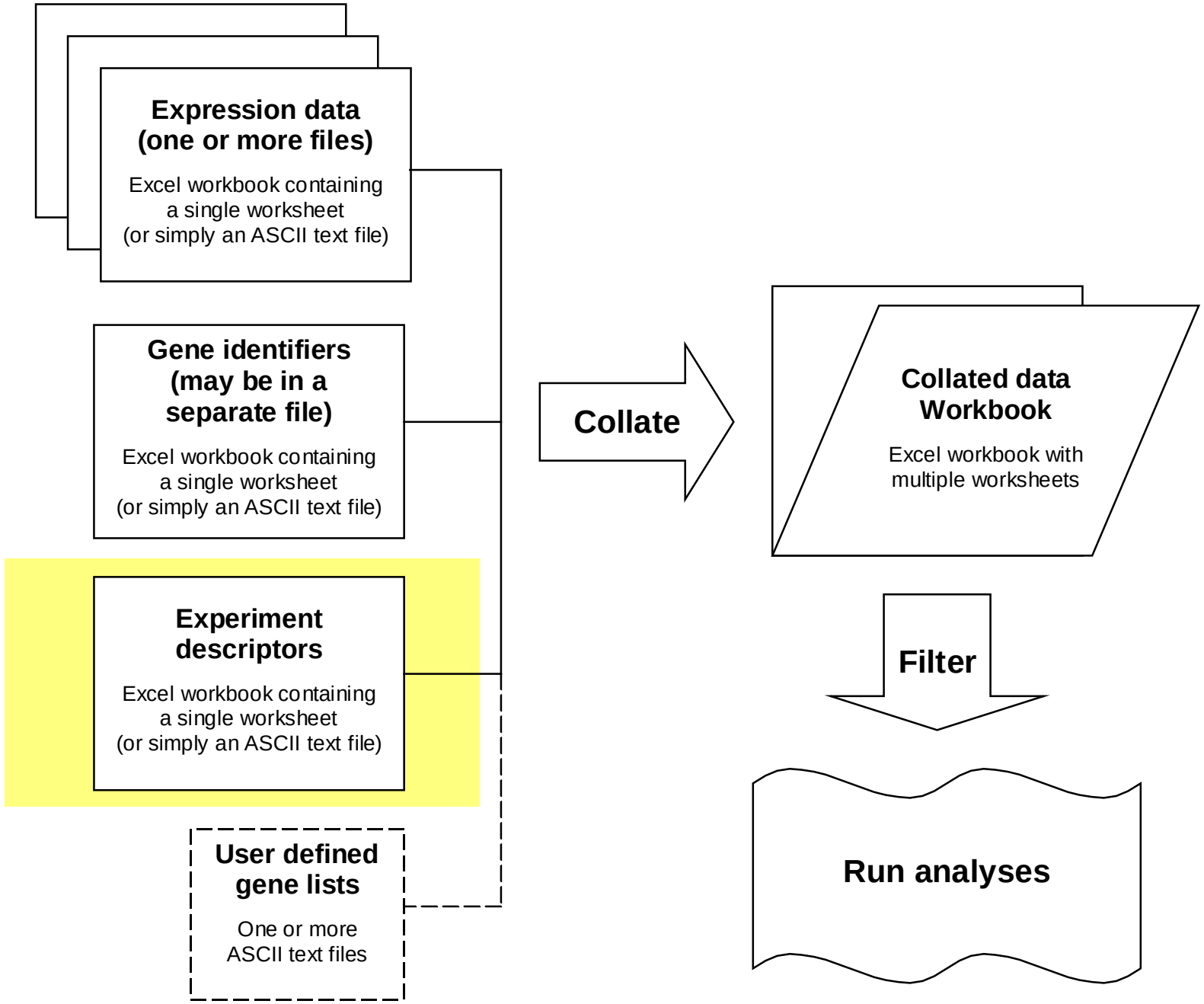
| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Spot | Clone | Description | GB acc | |
| 2 | 49 | 60204 | Homo sapiens C2H2 zinc finger protein pseudogene, mRNA sequence | T39154, T40438 | |
| 3 | 50 | 60436 | RPL3 Ribosomal protein L3 Chr.22 | T39295, T40510 | |
| 4 | 51 | 60218 | ESTs | T39165, T40450 | |
| 5 | 52 | 60209 | ESTs | T39163, T40448 | |
| 6 | 53 | 60664 | ESTs | T39448, T40595 | |
| 7 | 54 | 60932 | CSH1 Chorionic somatomammotropin hormone 1 (placental lactogen) Chr.17 | T39603, T40692 | |

GeneIds

Gene_identifiers.xls

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Well_id | Clone | Description | UniGene | Gene | Map | |
| 2 | 16027 | IMAGE:809353 | IRF-3=interferon regulatory factor-3 | Hs.75254 | IRF3 | 19q13.3-q13.4 | |
| 3 | 16028 | IMAGE:668442 | Receptor protein tyrosine kinase TKT precursor=Tyrosi | Hs.71891 | DDR2 | 1q12-q23 | |
| 4 | 16029 | IMAGE:767183 | HS1= hematopoietic lineage cell specific protein = hom | Hs.14601 | HCLS1 | 3q13 | |
| 5 | 4620 | IMAGE:485857 | delta sleep inducing peptide, immunoreactor | Hs.75450 | DSIPI | Xp21.1-q25 | |
| 6 | 4621 | IMAGE:485882 | P-selectin glycoprotein ligand | Hs.79283 | SELPLG | 12q24 | |
| 7 | 4622 | IMAGE:486003 | mrg1=melanocyte-specific nuclear protein associated w | Hs.82071 | CITED2 | 6q23.3 | |
| 8 | 4623 | IMAGE:485885 | CREG=cellular repressor of E1A-stimulated genes | Hs.5710 | CREG | 1q24 | |
| 9 | 4624 | IMAGE:485770 | Tis11d=ERF-2=growth factor early response gene | Hs.78909 | BRF2 | 2p22.3-2p21 | |

Gene_identifiers

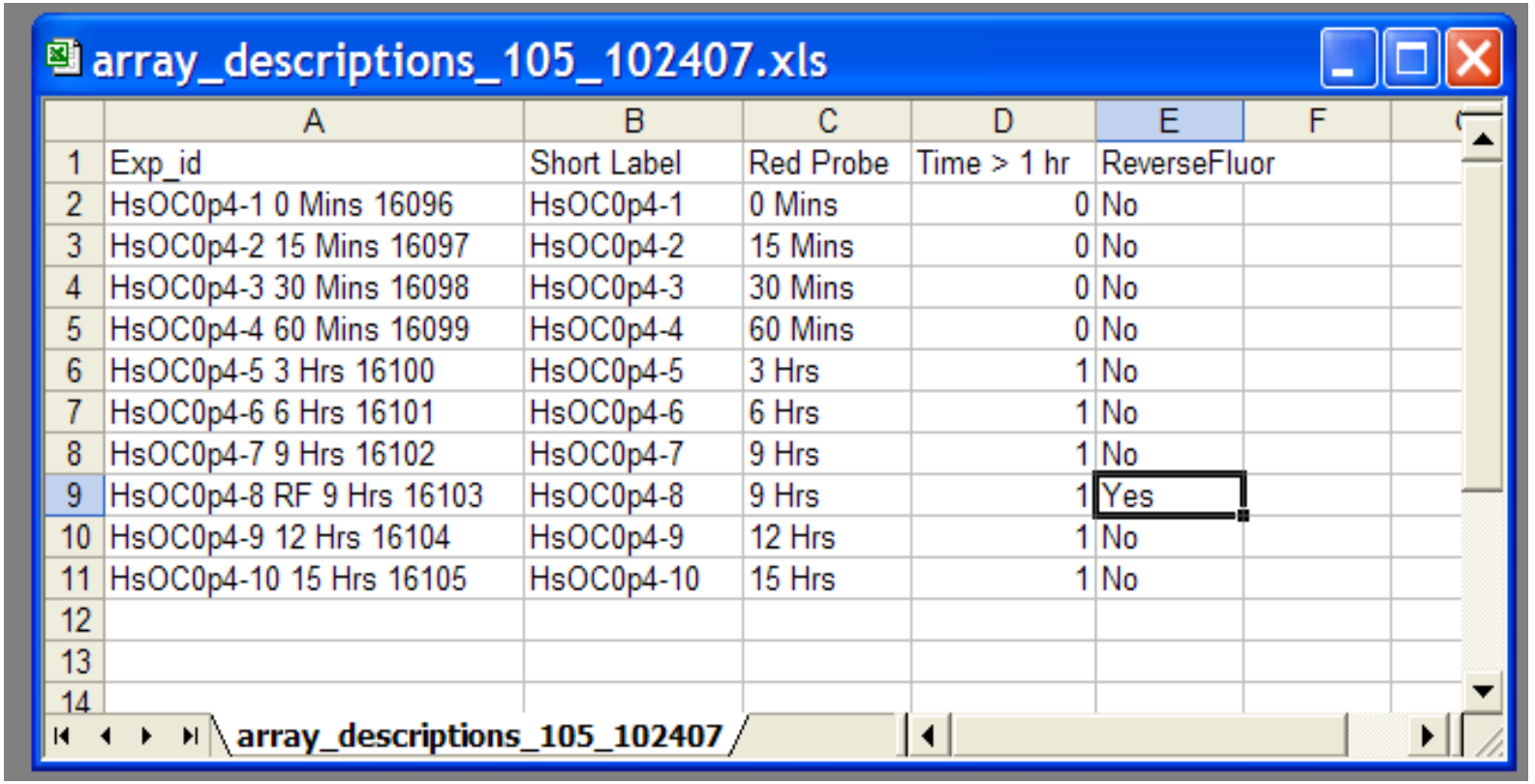

Expression data
(one or more files)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Gene identifiers
(may be in a
separate file)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Experiment
descriptors
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
User defined
gene lists
One or more
ASCII text files
Collate
Collated data
Workbook
Excel workbook with
multiple worksheets
Filter
Run analyses

# Experiment (Array) descriptors

- An experiment descriptors file describes the samples used for each array, and is mandatory.
- After the header row, each row in this file represents one array or sample, and each column represents one descriptor variable.
- First column contains array id, which is matched against file names when expression data is in separate files format.
- Subsequent columns contain descriptions, phenotype class labels, patient outcome, and other sample or experiment information.
- The descriptor variable columns may include information such as: patient ids, class labels, technical replicate indicators, reverse fluor indicators, and other variables used for labeling purposes.
- A COPY of the original experiment descriptor file will appear in the experiment descriptor sheet of the collated project workbook. The experiment descriptor sheet in the collated project workbook may be further edited as you analyze the data.

# Experiment descriptors

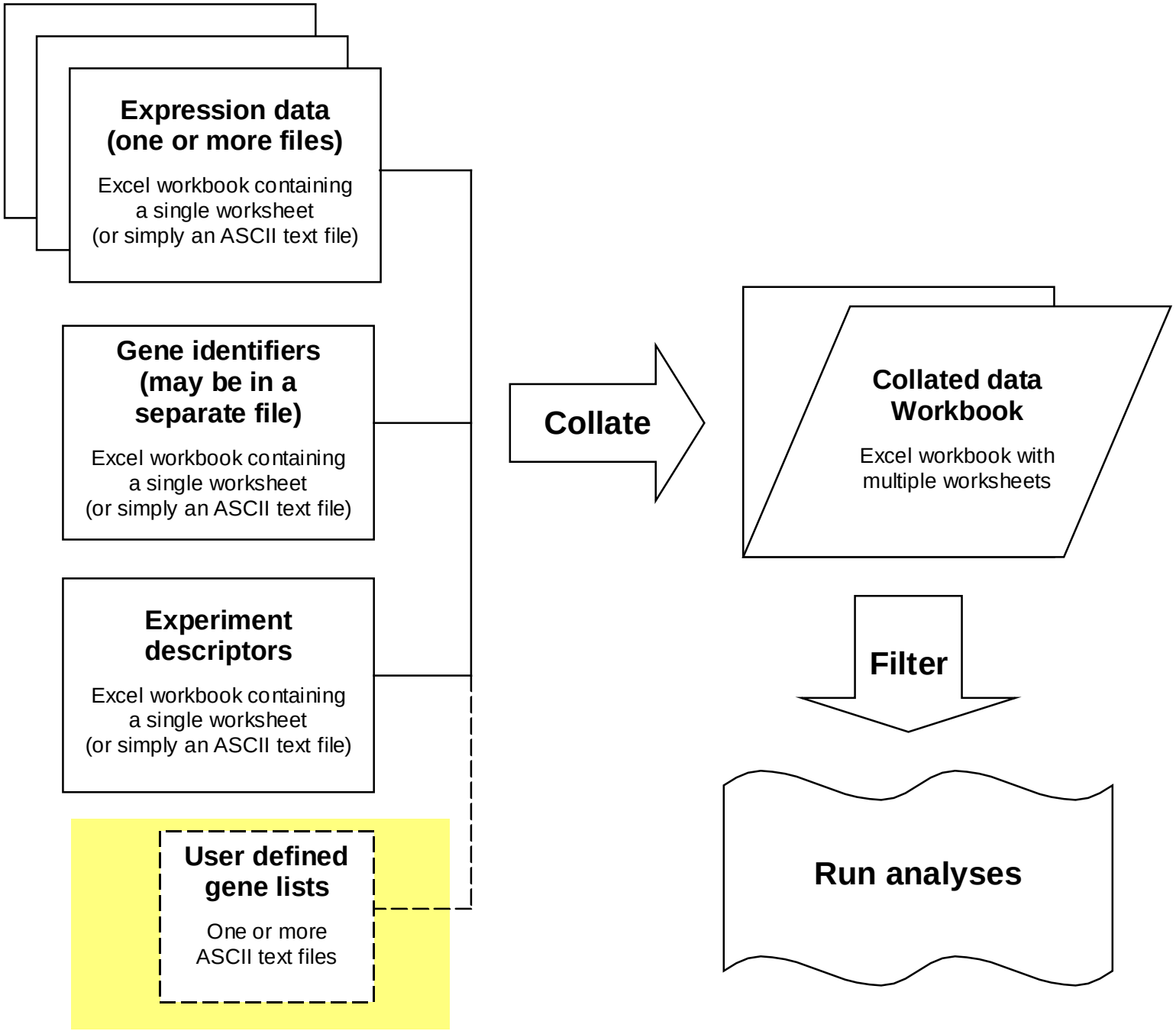
Describes the samples used for each array



array_descriptions_105_102407.xls

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Exp_id | Short Label | Red Probe | Time > 1 hr | ReverseFluor | |
| 2 | HsOC0p4-1 0 Mins 16096 | HsOC0p4-1 | 0 Mins | 0 | No | |
| 3 | HsOC0p4-2 15 Mins 16097 | HsOC0p4-2 | 15 Mins | 0 | No | |
| 4 | HsOC0p4-3 30 Mins 16098 | HsOC0p4-3 | 30 Mins | 0 | No | |
| 5 | HsOC0p4-4 60 Mins 16099 | HsOC0p4-4 | 60 Mins | 0 | No | |
| 6 | HsOC0p4-5 3 Hrs 16100 | HsOC0p4-5 | 3 Hrs | 1 | No | |
| 7 | HsOC0p4-6 6 Hrs 16101 | HsOC0p4-6 | 6 Hrs | 1 | No | |
| 8 | HsOC0p4-7 9 Hrs 16102 | HsOC0p4-7 | 9 Hrs | 1 | No | |
| 9 | HsOC0p4-8 RF 9 Hrs 16103 | HsOC0p4-8 | 9 Hrs | 1 | Yes | |
| 10 | HsOC0p4-9 12 Hrs 16104 | HsOC0p4-9 | 12 Hrs | 1 | No | |
| 11 | HsOC0p4-10 15 Hrs 16105 | HsOC0p4-10 | 15 Hrs | 1 | No | |
| 12 | | | | | | |
| 13 | | | | | | |
| 14 | | | | | | |

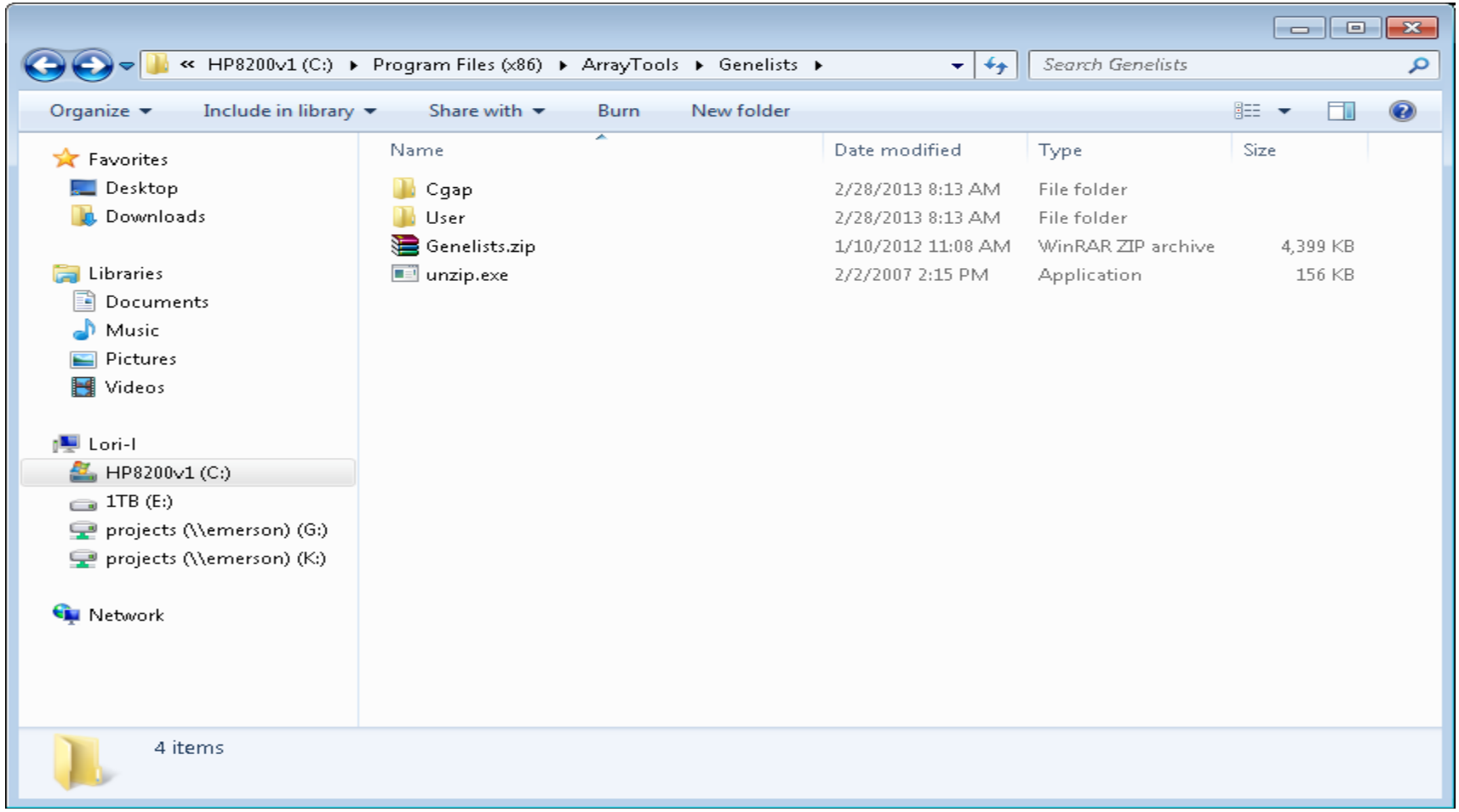
array_descriptions_105_102407


Expression data
(one or more files)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Gene identifiers
(may be in a
separate file)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Experiment
descriptors
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
User defined
gene lists
One or more
ASCII text files
Collate
Collated data
Workbook
Excel workbook with
multiple worksheets
Filter
Run analyses

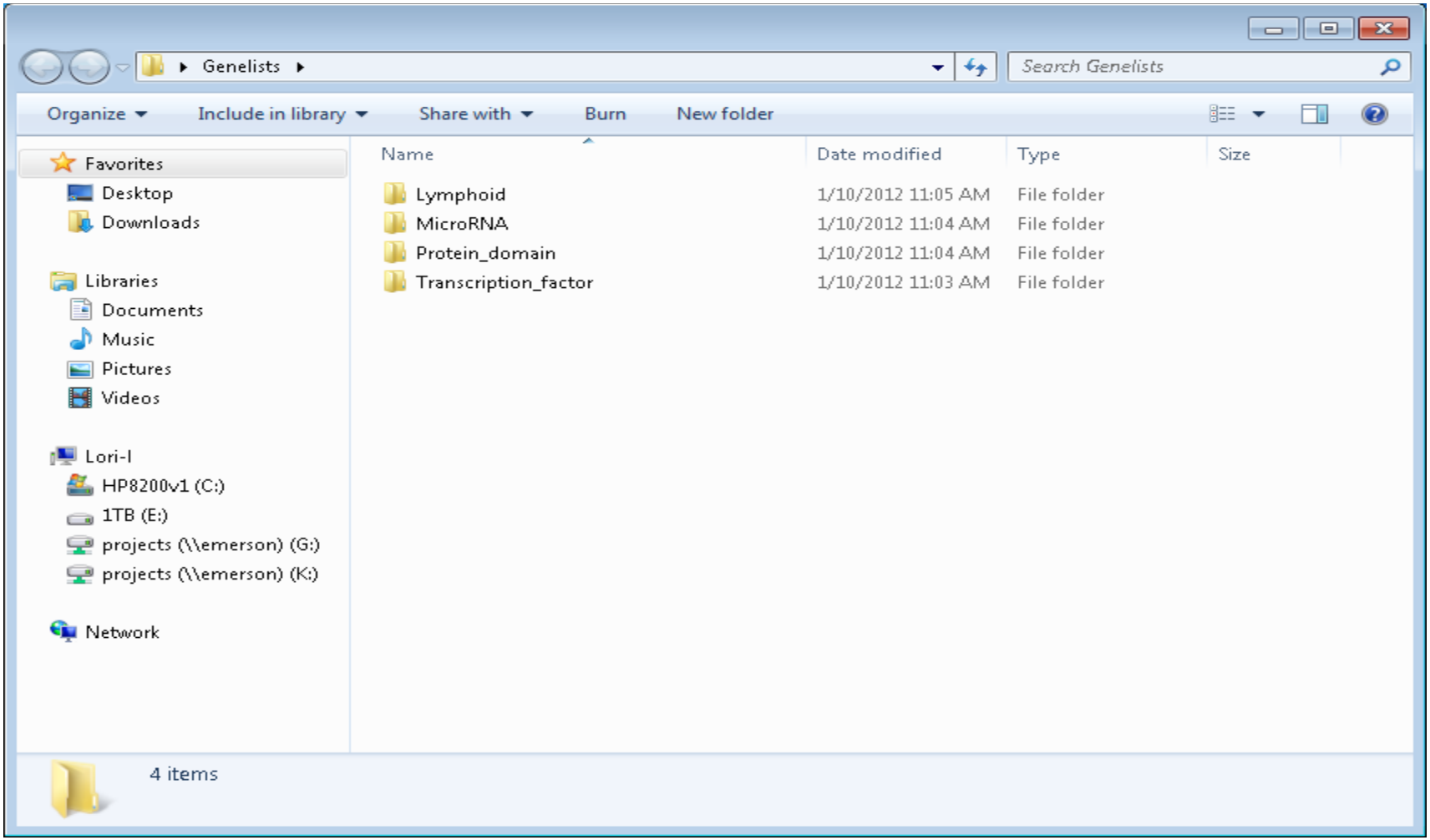
# Gene lists

- Genelists are used for annotation and for defining subsets for analysis. These files are located in the ArrayTools installation folder.
- Two types of genelists:  Preloaded with BRB-ArrayTools, and user-defined
- User-defined genelists are simply text files which the user creates, containing a label specifying the type of identifier, followed by a list of gene identifiers.  The file should be appropriately named to indicate what type of genes are in the list.  Some user-defined genelists are automatically produced as the result of an analysis, such as class comparison, class prediction, survival analysis, and hierarchical clustering of genes.
- User-defined genelists are stored in the “project” folder (for project specific) or ArrayTools folder (visible to  all projects.)

# Genelists pre-loaded with BRB-ArrayTools

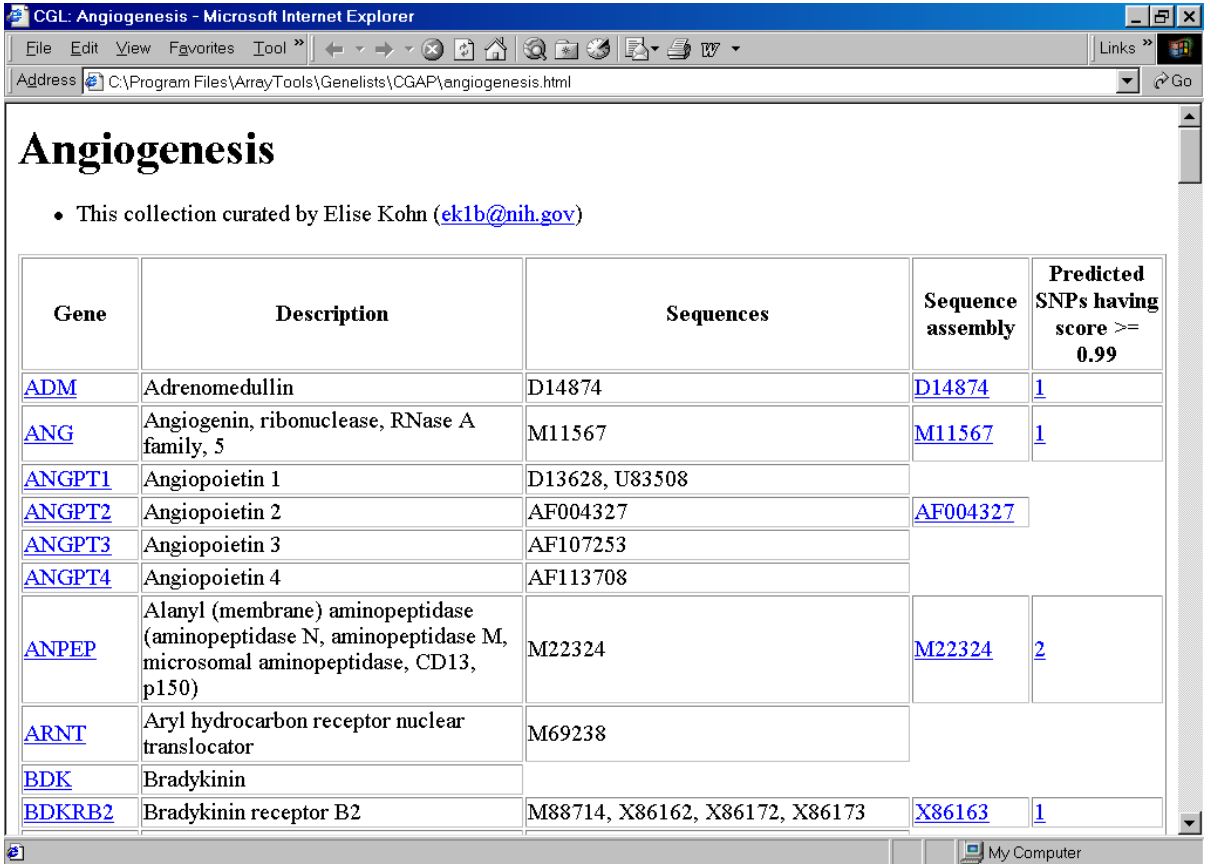


# Genelists Folder



# Gene lists

## Cancer Genome Anatomy Project




| Gene | Description | Sequences | Sequence assembly | Predicted SNPs having score >= 0.99 |
|---|---|---|---|---|
| ADM | Adrenomedullin | D14874 | D14874 | 1 |
| ANG | Angiogenin, ribonuclease, RNase A family, 5 | M11567 | M11567 | 1 |
| ANGPT1 | Angiopoietin 1 | D13628, U83508 | | |
| ANGPT2 | Angiopoietin 2 | AF004327 | AF004327 | |
| ANGPT3 | Angiopoietin 3 | AF107253 | | |
| ANGPT4 | Angiopoietin 4 | AF113708 | | |
| ANPEP | Alanyl (membrane) aminopeptidase (aminopeptidase N, aminopeptidase M, microsomal aminopeptidase, CD13, p150) | M22324 | M22324 | 2 |
| ARNT | Aryl hydrocarbon receptor nuclear translocator | M69238 | | |
| BDK | Bradykinin | | | |
| BDKRB2 | Bradykinin receptor B2 | M88714, X86162, X86172, X86173 | X86163 | 1 |

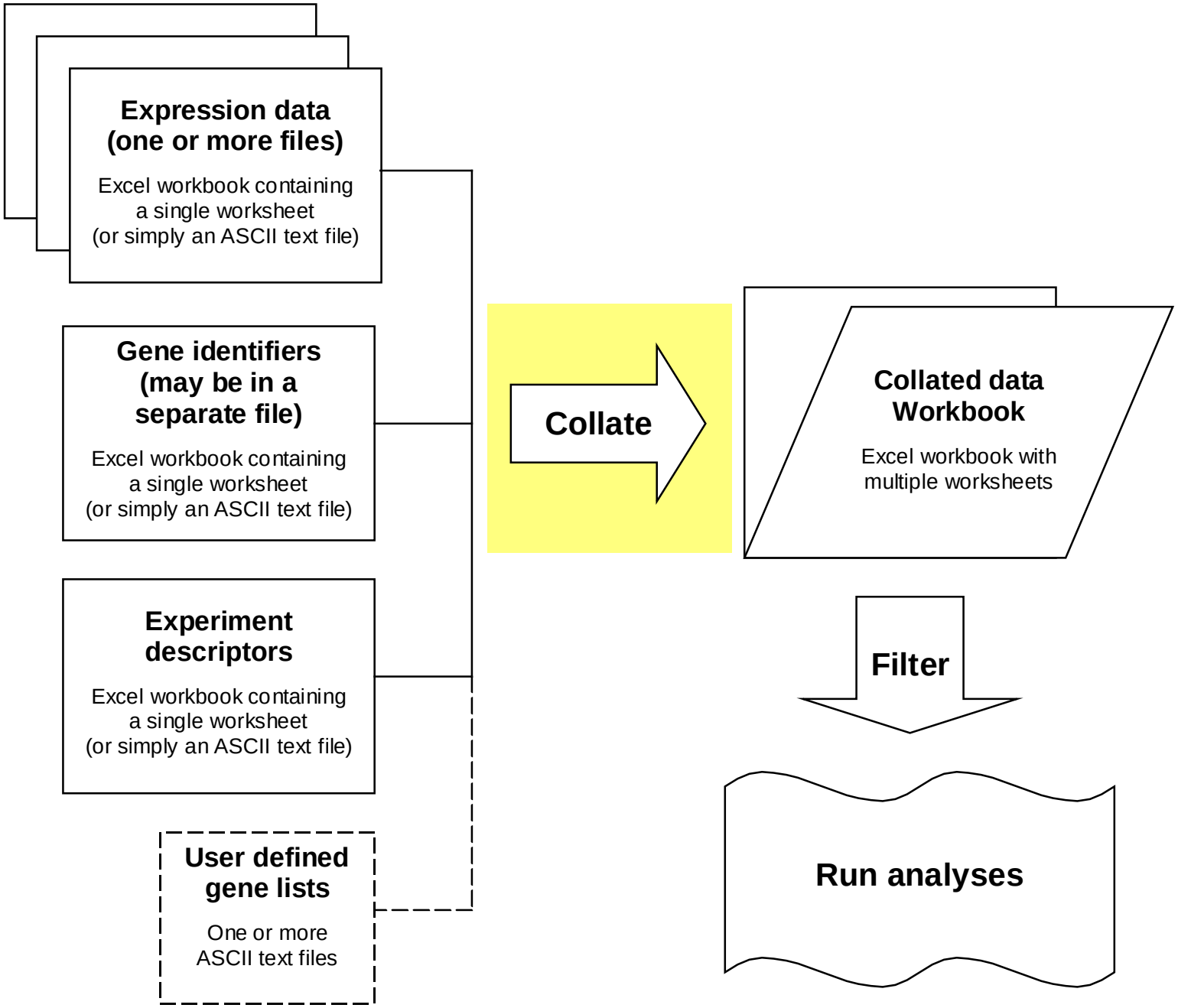
# Gene lists

## User-defined text files

ClassComparison.txt - Notepad

File Edit Format View Help

| ProbeSet | Accession | UGCluster | Symbol | EntrezID |
|---|---|---|---|---|
| S82024_at | S82024 | Hs.521651 | STMN2 | 11075 |
| U52828_s_at | U52828 | Hs.314543 | CTNND2 | 1501 |
| D21267_at | D21267 | Hs.167317 | SNAP25 | 6616 |
| M14483_rna1_s_at | M14483 | Hs.459927 | PTMA | 5757 |
| U38810_at | U38810 | Hs.584776 | MAB21L1 | 4081 |
| D82347_at | D82347 | Hs.574626 | NEUROD1 | 4760 |
| M93119_at | M93119 | Hs.89584 | INSM1 | 3642 |
| X86809_at | X86809 | Hs.517216 | PEA15 | 8682 |
| L10373_at | L10373 | Hs.441664 | TSPAN7 | 7102 |
| U50822_rna1_s_at | U50822 | Hs.574626 | NEUROD1 | 4760 |
| U76421_at | U76421 | Hs.474018 | ADARB1 | 104 |
| U29195_at | U29195 | Hs.3281 NPTX2 | 4885 | |
| U00802_s_at | U00802 | Hs.130316 | DBN1 | 1627 |
| U96136_at | U96136 | Hs.314543 | CTNND2 | 1501 |
| M33308_at | M33308 | Hs.643896 | VCL | 7414 |
| X02761_s_at | X02761 | Hs.203717 | FN1 | 2335 |
| HG3454-HT3647_at | HG3454-HT3647 | | | |
| D49958_at | D49958 | Hs.75819 | GPM6A | 2823 |
| S45630_at | S45630 | Hs.53454 | CRYAB | 1410 |
| M25667_at | M25667 | Hs.134974 | GAP43 | 2596 |
| L48513_at | L48513 | Hs.530077 | PON2 | 5445 |
| M29536_at | M29536 | Hs.429180 | EIF2S2 | 8894 |
| M97287_at | M97287 | Hs.517717 | SATB1 | 6304 |
| X70476_at | X70476 | Hs.75724 | COPB2 | 9276 |
| U25034_s_at | U25034 | Hs.504703 | NNAT | 4826 |
| L10338_s_at | L10338 | Hs.436646 | SCN1B | 6324 |
| X51405_at | X51405 | Hs.75360 | CPE | 1363 |
| S78296_at | S78296 | Hs.500916 | INA | 9118 |
| M98539_at | M98539 | Hs.446429 | PTGDS | 5730 |
| U30521_at | U30521 | Hs.36053 | C5orf13 | 9315 |
| M37457_at | M37457 | Hs.515427 | ATP1A3 | 478 |
| AJ001421_at | AJ001421 | Hs.525527 | RER1 | 11079 |
| X05196_at | X05196 | Hs.155247 | ALDOC | 230 |
| HG658-HT658_f_at | HG658-HT658 | | | |
| HG3236-HT3413_f_at | HG3236-HT3413 | | | |
| U45955_at | U45955 | Hs.495710 | GPM6B | 2824 |
| M94250_at | M94250 | Hs.82045 | MDK | 4192 |
| M93426_at | M93426 | Hs.489824 | PTPRZ1 | 5803 |
| HG3597-HT3800_f_at | HG3597-HT3800 | | | |
| U90915_at | U90915 | Hs.433419 | COX4I1 | 1327 |
| U60975_at | U60975 | Hs.368592 | SORL1 | 6653 |
| Y09836_at | Y09836 | Hs.335079 | MAP1B | 4131 |
| S54005_s_at | S54005 | Hs.446574 | TMSB10 | 9168 |
| D12676_at | D12676 | Hs.349656 | SCARB2 | 950 |
| D78012_at | D78012 | Hs.135270 | CRMP1 | 1400 |
| U48705_rna1_s_at | U48705 | Hs.631988 | DDR1 | 780 |
| U76638_at | U76638 | Hs.591642 | BARD1 | 580 |
| J04173_at | J04173 | Hs.632918 | PGAM1 | 5223 |


Expression data
(one or more files)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Gene identifiers
(may be in a
separate file)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Experiment
descriptors
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
User defined
gene lists
One or more
ASCII text files
Collate
Collated data
Workbook
Excel workbook with
multiple worksheets
Filter
Run analyses

# Specify data using the collate dialog form

- Expression data: Specify the expression data file (or folder), and data columns within the data file(s)
- Gene identifiers: Specify the file, and columns containing the identifiers (specify hyperlinkable gene identifiers separately)
- Experiment descriptors: Specify the file, and reverse fluor indicators (if any)

# Automatic data importers

- General format data: The general format importer can be used to import most data formats.
- Data Import Wizard: This importer can be used to import a variety of data types.
- NCBI GEO archive Importer: This importer can be used to import GDS datasets from GEO archive.
- Affymetrix gene 1.0 st array importer.

# Data types

- The following data types can be imported through Data Import Wizard:
  - 3’-IVT chip data in .CEL file format
  - Affymetrix gene 1.0 ST array data in .CEL file format
  - Affymetrix probe-set summary level data
  - Illumina expression data
  - Illumina methylation data
  - AgilentAffymetrix  single and dual-channel data
  - Genepix single and dual-channel data
  - mAdb archive data
  - RNA-Seq data pre-processed by Galaxy

# Part III:

# The collated project workbook

# Pomeroy Dataset

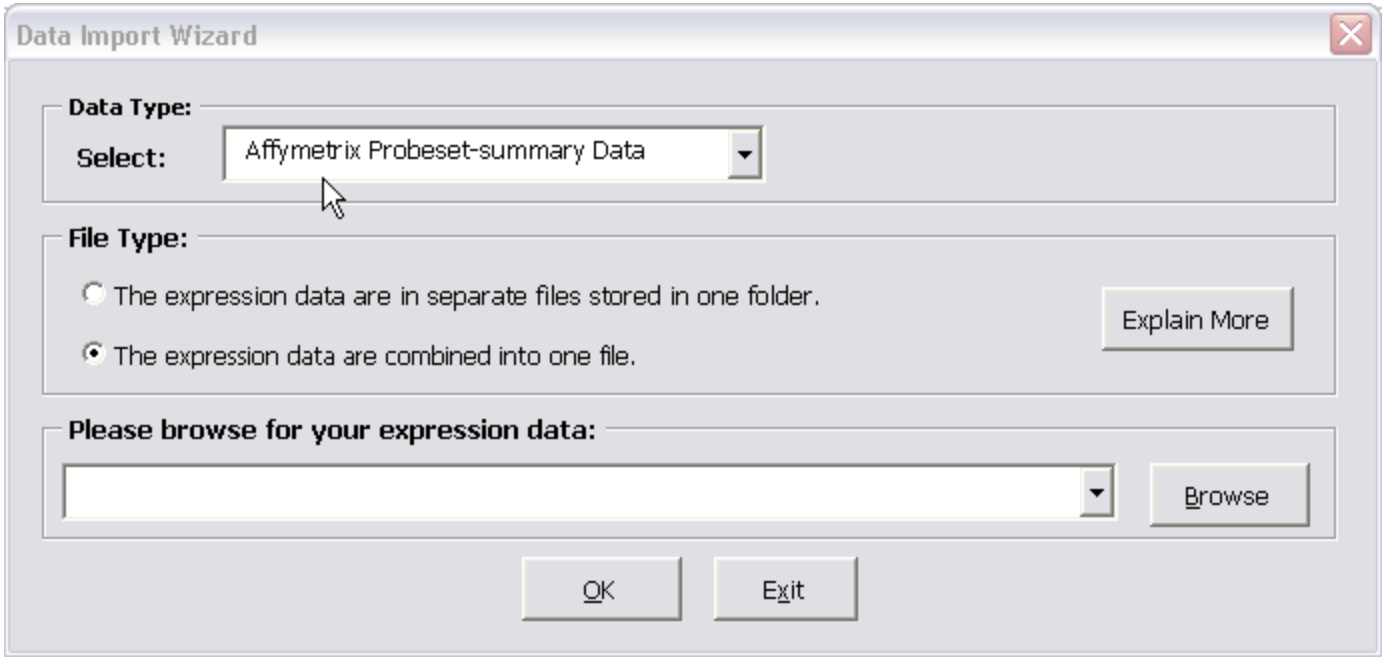
- **On the Desktop**, browse for the folder called " **BRB-ArrayTools-Class**".
- Under this folder, look for the sub-folder "**Pomeroy**".
- In this folder there are two files namely:

  **Dataset_A2_multiple_tumor_samples.txt**
  **ExpDescrMedulo.xls**
- The **Dataset_A2_multiple_tumor samples.txt** contains the raw expression MAS5.0 summary values for all the arrays.
- The **ExpDescrMedulo.xls** contains the experiment descriptor file.

# [Hands-on instructions]

## [Importing Pomeroy Data set]

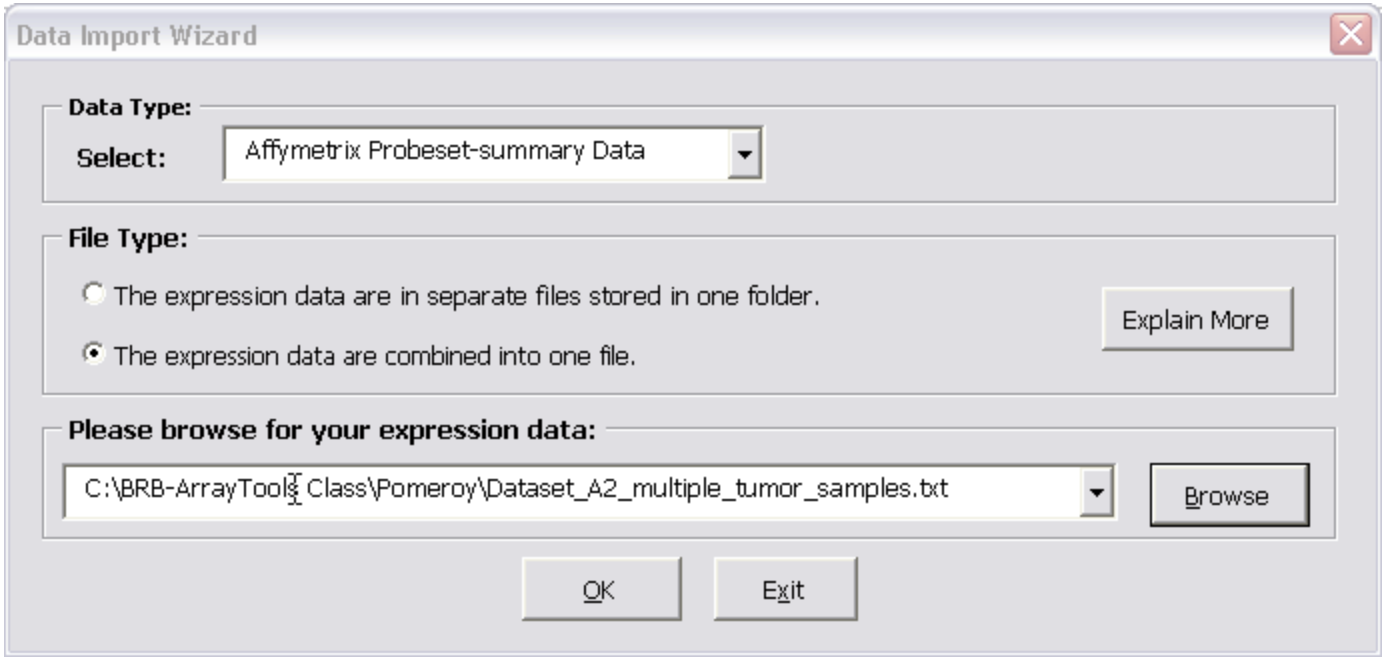
- Click on **ArrayTools →Getting started → Data Import Wizard**
- Select the option from the pull down menu- **"Affymetrix probeset-summary data".**
- **Choose the option that the expression data is combined into one file.**



# [Hands-on instructions]

## [Importing Pomeroy Data set]

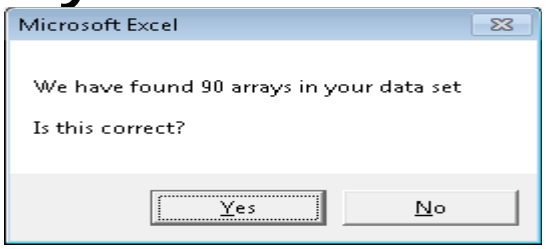
•**Browse** for the following file which is also in the **Pomeroy** folder inside the **BRB-ArrayTools Class** folder which is on the **Desktop**: **`Dataset_A2_multiple_tumor_samples.txt`** and then click **OK.**



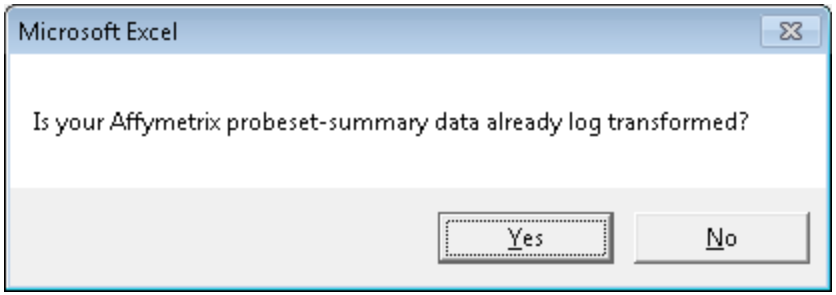
# [Hands-on instructions]

## [Importing Pomeroy Data set]

- Click “yes” to the following question on number of arrays.



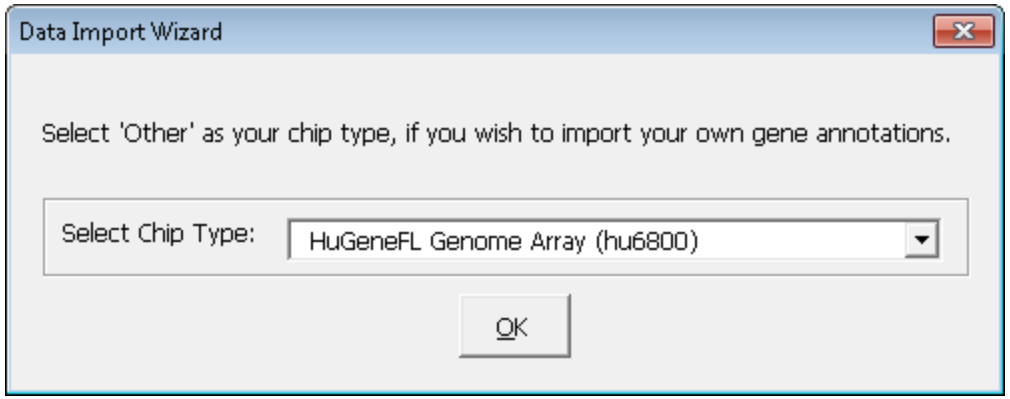

- Click “No” to the question about log transformation.



# [Hands-on instructions]

## [Importing Pomeroy Data set]

- Select the chip type as “HuGeneFL Genome Array”

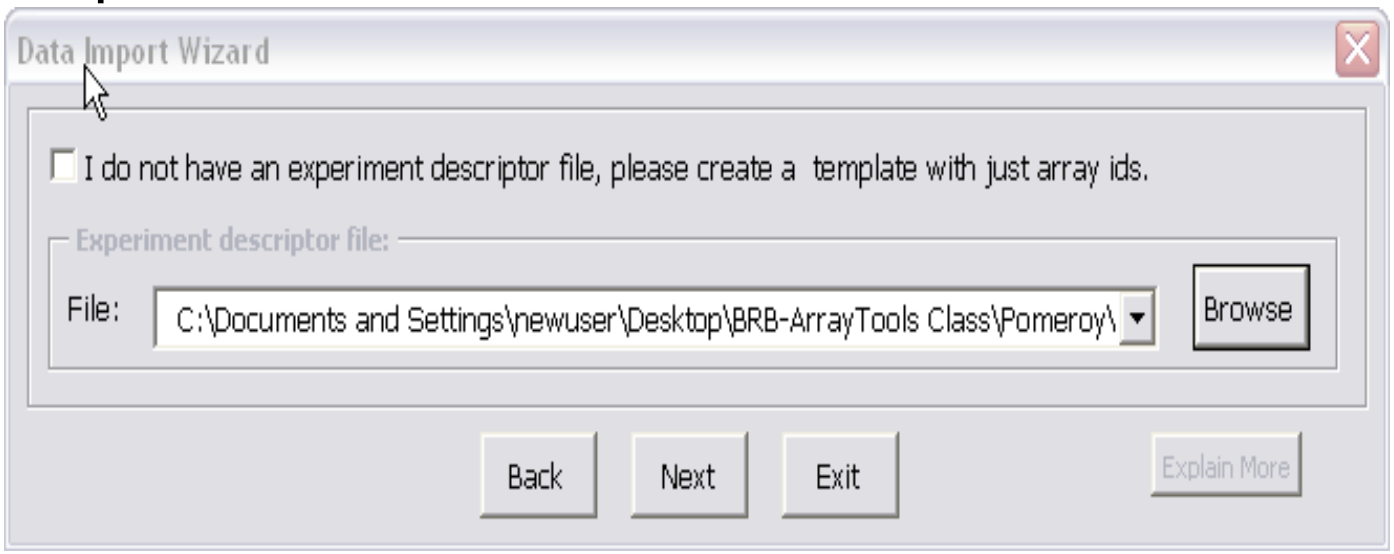


# [Hands-on instructions]

## [Importing Pomeroy Data set]

- **Browse** for the following file in the **`Pomeroy`** folder inside the **`BRB-ArrayTools Class`** folder which is on the **Desktop**:

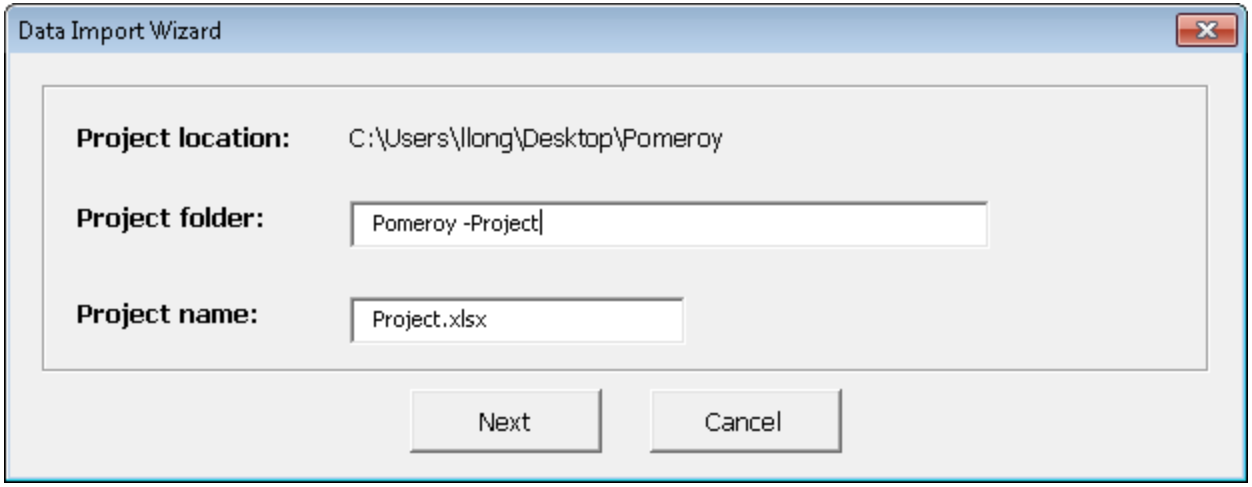
  "ExpDescrMedulo.xls" and click "Next".



# [Hands-on instructions]

## [Importing Pomeroy Data set]

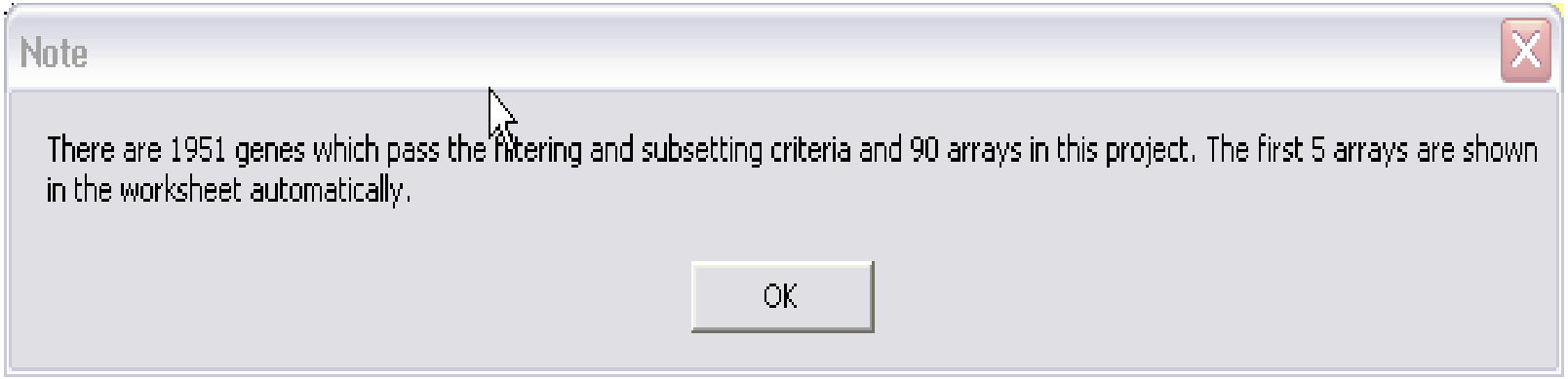
- Save the Project in the folder “Pomeroy-Project”.



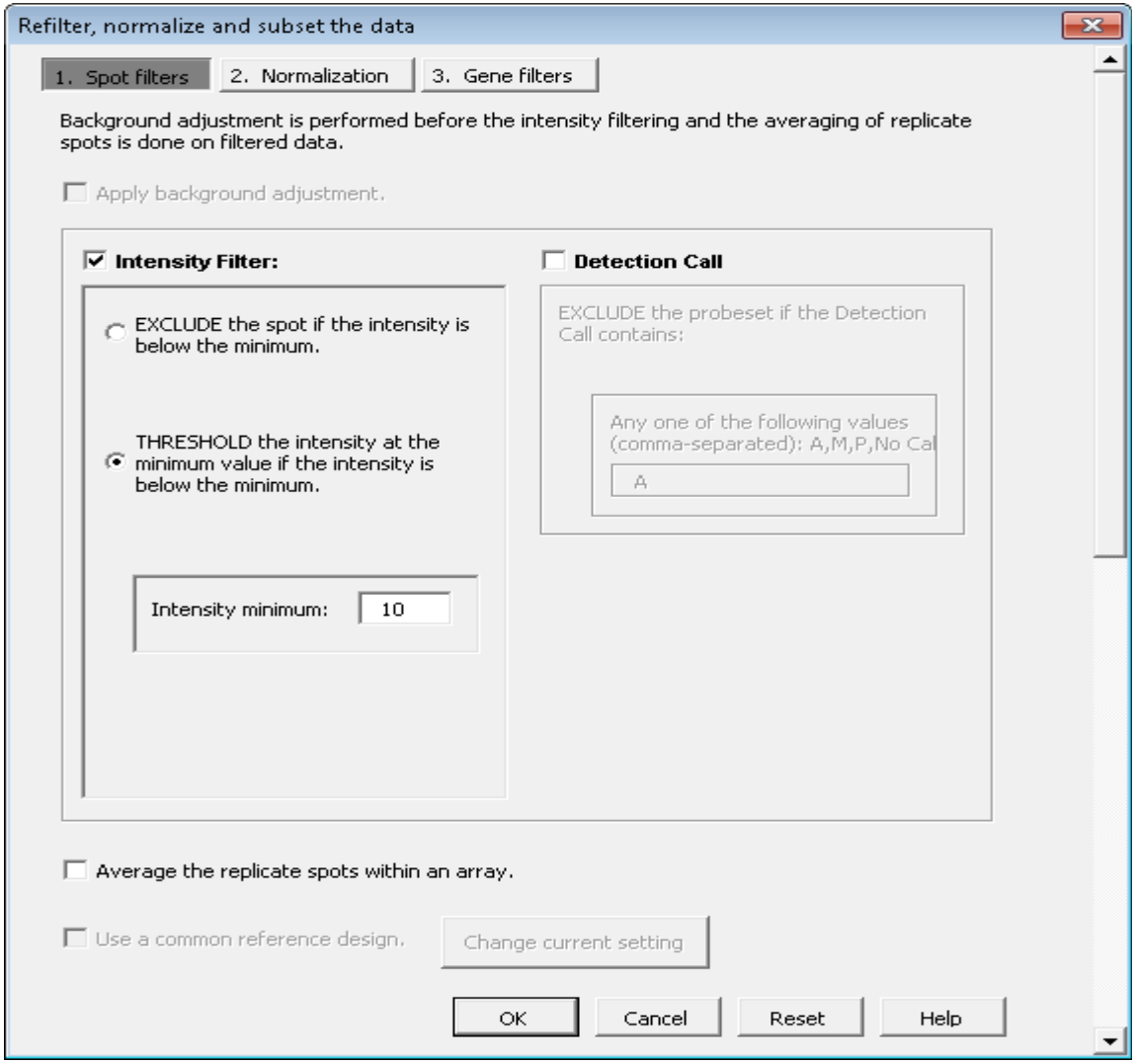
# [Hands-on instructions]

## [Importing Pomeroy Data set]

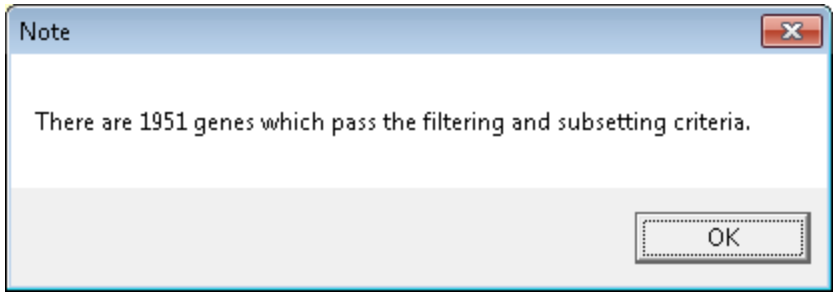
- The progress bar will indicate that the project is collating.
- Click “OK”



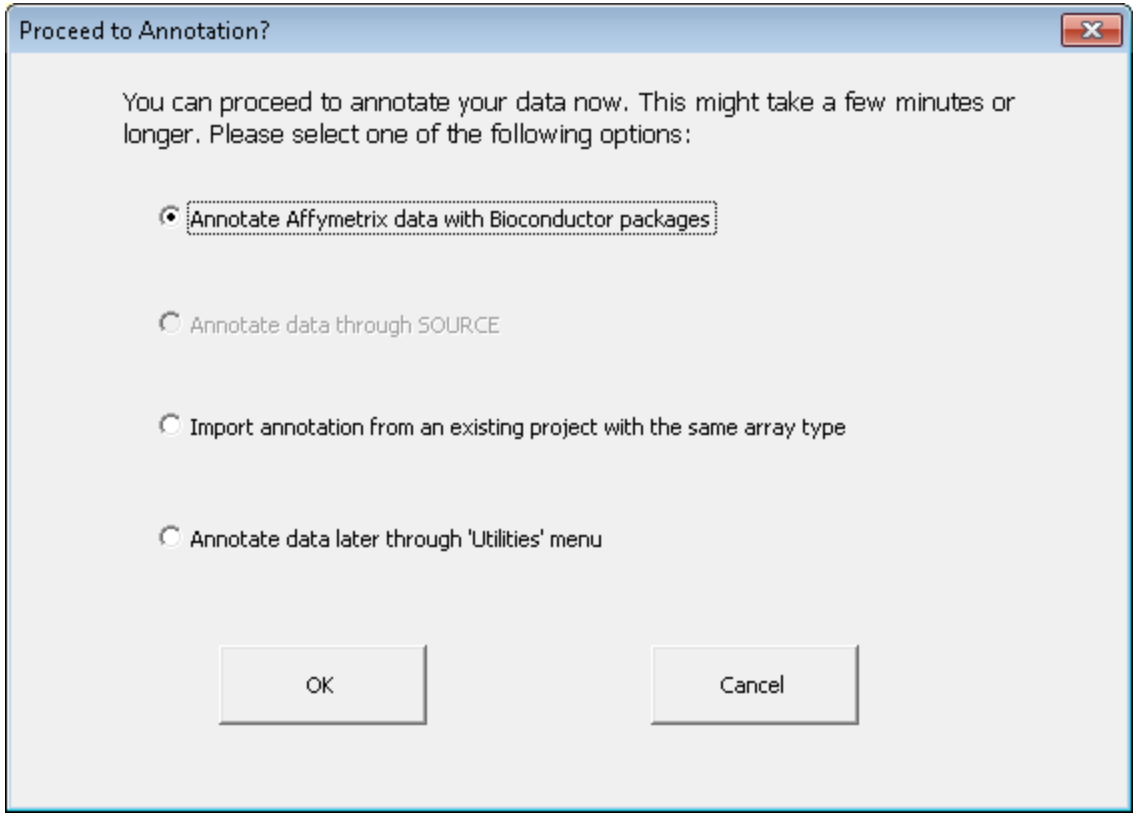
# Filtering and Normalization



# When importing is completed



# Annotation


Proceed to Annotation?
You can proceed to annotate your data now. This might take a few minutes or longer. Please select one of the following options:
Annotate Affymetrix data with Bioconductor packages
Annotate data through SOURCE
Import annotation from an existing project with the same array type
Annotate data later through 'Utilities' menu
OK
Cancel

# Collated project workbook

## Overview

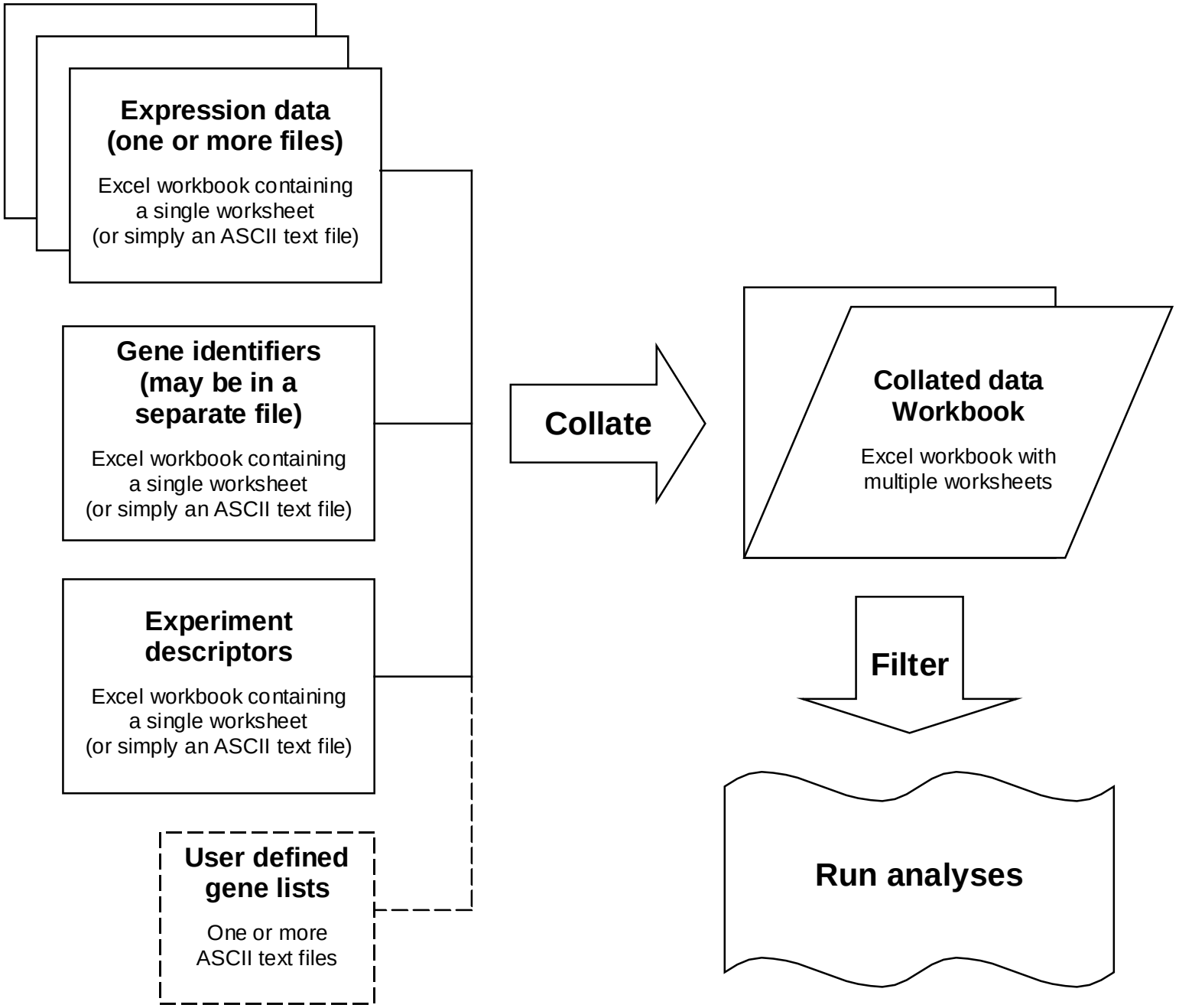
- The **collated project workbook** is the primary data object on which future analyses are run
- The collated project workbook is located inside the **project folder**, which by default is located inside the folder where the original input data is located.
- The project folder may also contain some other folders: **`BinaryData`**, **`Annotations`**, **`Output`**, and **`Genelists`**.
- The **`BinaryData`** and **`Annotations`** folders should NOT be altered by users. These are used for internal purposes.
- The **`Output`** folder will contain the output of all subsequent analyses.
- A **`Genelists`** folder may also be created, and may contain genelists to be used for subset analyses.

# The collated project workbook

- This is the primary data object on which future analyses are run.

- Contains three primary worksheets:
  1. Experiment descriptors (may edit this to specify analyses)
  2. Gene identifiers
  3. Filtered log ratio  (or Filtered log intensity)

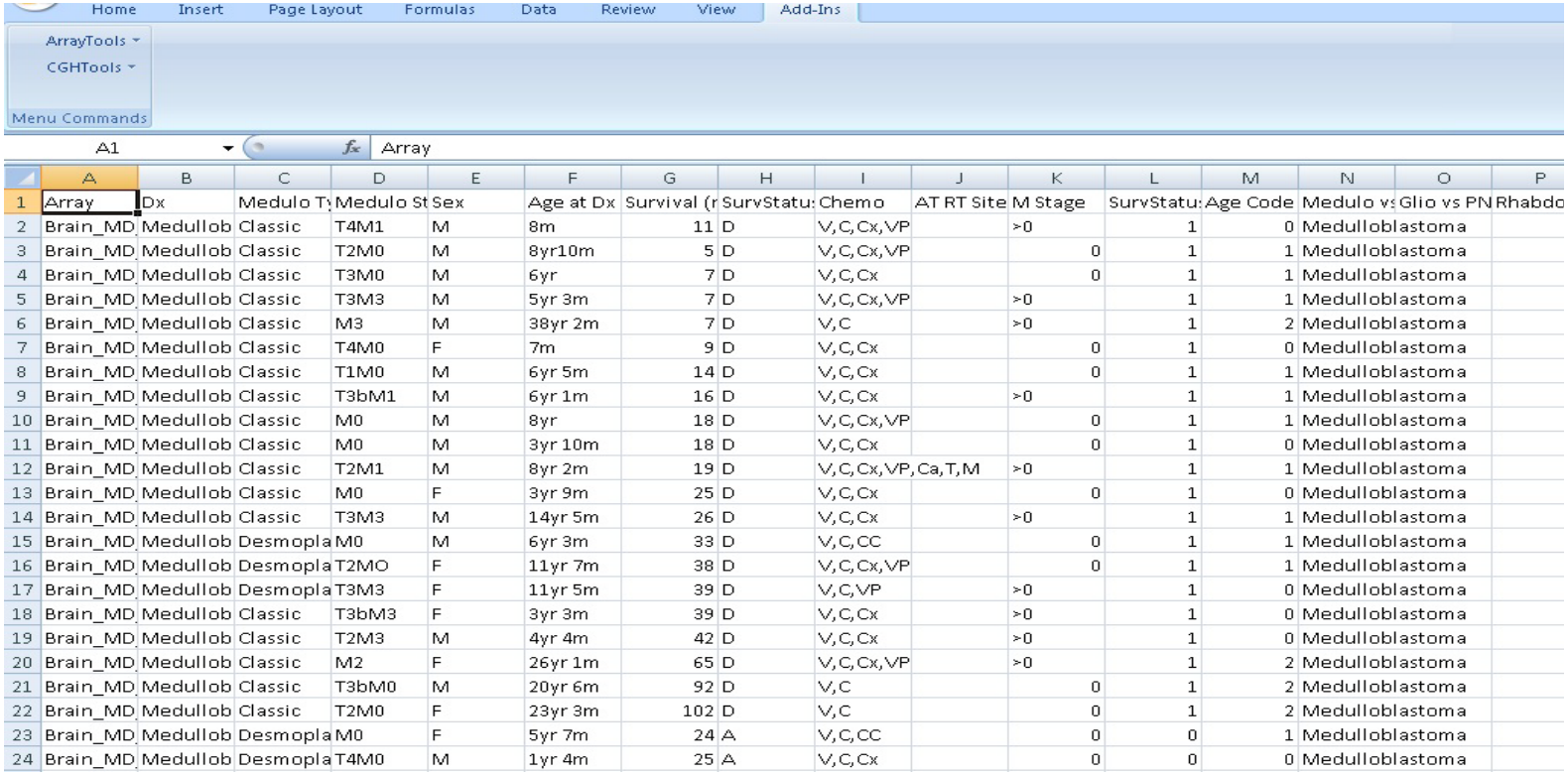
- Additional results worksheets which may be automatically added:
  1. Gene annotations (obtained by running the menu item:
     **Utilities → Annotate data →**
     **Import Affymetrix or SOURCE annotations**)
  2. Scatterplot results
  3. Cluster analysis results


Expression data
(one or more files)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Gene identifiers
(may be in a
separate file)
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
Experiment
descriptors
Excel workbook containing
a single worksheet
(or simply an ASCII text file)
User defined
gene lists
One or more
ASCII text files
Collate
Collated data
Workbook
Excel workbook with
multiple worksheets
Filter
Run analyses

# The collated project workbook

## Experiment descriptor sheet

Create experiment descriptor variables which can be used to guide and specify the analyses.

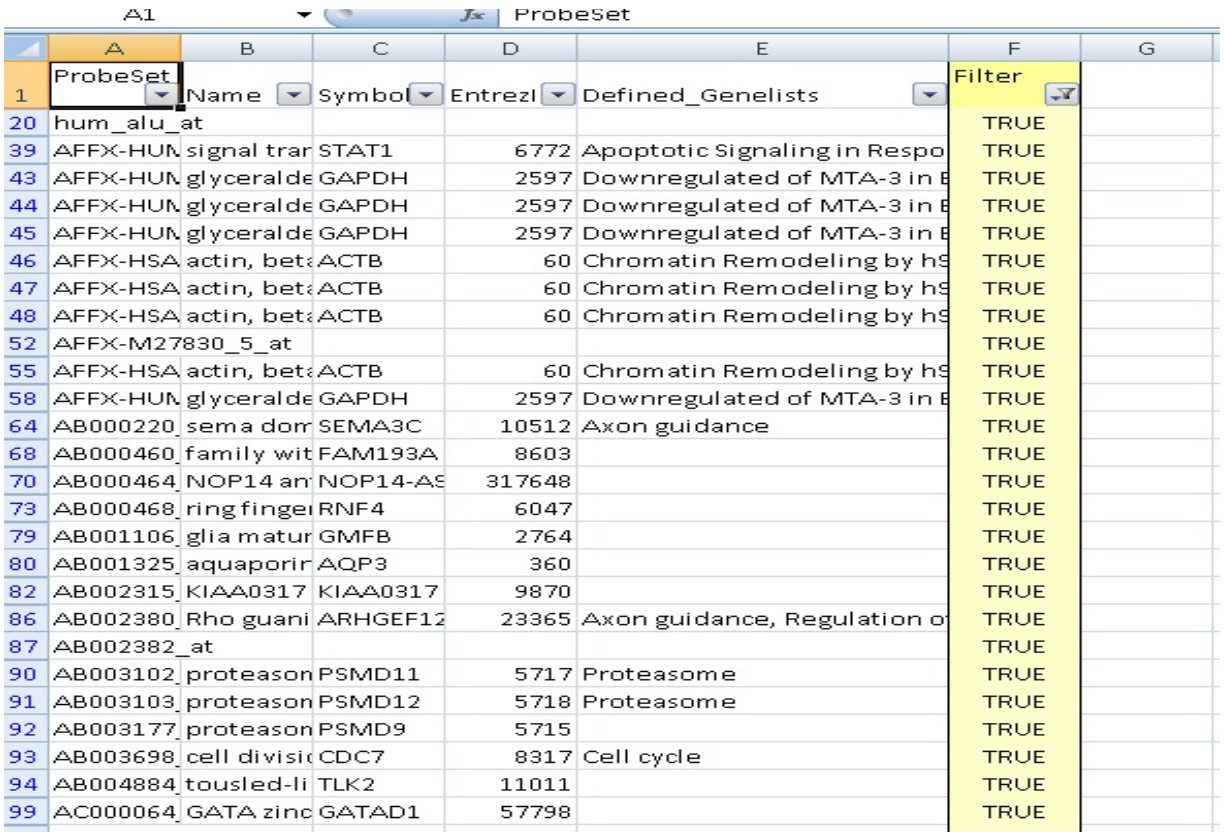



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Array | Dx | Medulo Ty | Medulo St | Sex | Age at Dx | Survival (r | SurvStatus | Chemo | AT RT Site | M Stage | SurvStatus | Age Code | Medulo vs | Glio vs PN | Rhabdo |
| 2 | Brain_MD | Medullob | Classic | T4M1 | M | 8m | 11 | D | V,C,Cx,VP | | >0 | 1 | 0 | Medulloblastoma | | |
| 3 | Brain_MD | Medullob | Classic | T2M0 | M | 8yr10m | 5 | D | V,C,Cx,VP | | 0 | 1 | 1 | Medulloblastoma | | |
| 4 | Brain_MD | Medullob | Classic | T3M0 | M | 6yr | 7 | D | V,C,Cx | | 0 | 1 | 1 | Medulloblastoma | | |
| 5 | Brain_MD | Medullob | Classic | T3M3 | M | 5yr 3m | 7 | D | V,C,Cx,VP | | >0 | 1 | 1 | Medulloblastoma | | |
| 6 | Brain_MD | Medullob | Classic | M3 | M | 38yr 2m | 7 | D | V,C | | >0 | 1 | 2 | Medulloblastoma | | |
| 7 | Brain_MD | Medullob | Classic | T4M0 | F | 7m | 9 | D | V,C,Cx | | 0 | 1 | 0 | Medulloblastoma | | |
| 8 | Brain_MD | Medullob | Classic | T1M0 | M | 6yr 5m | 14 | D | V,C,Cx | | 0 | 1 | 1 | Medulloblastoma | | |
| 9 | Brain_MD | Medullob | Classic | T3bM1 | M | 6yr 1m | 16 | D | V,C,Cx | | >0 | 1 | 1 | Medulloblastoma | | |
| 10 | Brain_MD | Medullob | Classic | M0 | M | 8yr | 18 | D | V,C,Cx,VP | | 0 | 1 | 1 | Medulloblastoma | | |
| 11 | Brain_MD | Medullob | Classic | M0 | M | 3yr 10m | 18 | D | V,C,Cx | | 0 | 1 | 0 | Medulloblastoma | | |
| 12 | Brain_MD | Medullob | Classic | T2M1 | M | 8yr 2m | 19 | D | V,C,Cx,VP,Ca,T,M | | >0 | 1 | 1 | Medulloblastoma | | |
| 13 | Brain_MD | Medullob | Classic | M0 | F | 3yr 9m | 25 | D | V,C,Cx | | 0 | 1 | 0 | Medulloblastoma | | |
| 14 | Brain_MD | Medullob | Classic | T3M3 | M | 14yr 5m | 26 | D | V,C,Cx | | >0 | 1 | 1 | Medulloblastoma | | |
| 15 | Brain_MD | Medullob | Desmopla | M0 | M | 6yr 3m | 33 | D | V,C,CC | | 0 | 1 | 1 | Medulloblastoma | | |
| 16 | Brain_MD | Medullob | Desmopla | T2MO | F | 11yr 7m | 38 | D | V,C,Cx,VP | | 0 | 1 | 1 | Medulloblastoma | | |
| 17 | Brain_MD | Medullob | Desmopla | T3M3 | F | 11yr 5m | 39 | D | V,C,VP | | >0 | 1 | 0 | Medulloblastoma | | |
| 18 | Brain_MD | Medullob | Classic | T3bM3 | F | 3yr 3m | 39 | D | V,C,Cx | | >0 | 1 | 0 | Medulloblastoma | | |
| 19 | Brain_MD | Medullob | Classic | T2M3 | M | 4yr 4m | 42 | D | V,C,Cx | | >0 | 1 | 0 | Medulloblastoma | | |
| 20 | Brain_MD | Medullob | Classic | M2 | F | 26yr 1m | 65 | D | V,C,Cx,VP | | >0 | 1 | 2 | Medulloblastoma | | |
| 21 | Brain_MD | Medullob | Classic | T3bM0 | M | 20yr 6m | 92 | D | V,C | | 0 | 1 | 2 | Medulloblastoma | | |
| 22 | Brain_MD | Medullob | Classic | T2M0 | F | 23yr 3m | 102 | D | V,C | | 0 | 1 | 2 | Medulloblastoma | | |
| 23 | Brain_MD | Medullob | Desmopla | M0 | F | 5yr 7m | 24 | A | V,C,CC | | 0 | 0 | 1 | Medulloblastoma | | |
| 24 | Brain_MD | Medullob | Desmopla | T4M0 | M | 1yr 4m | 25 | A | V,C,Cx | | 0 | 0 | 0 | Medulloblastoma | | |

# The collated project workbook

## Gene identifier sheet

Contains gene identifiers provided by the user during collation.



A1 ProbeSet

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ProbeSet | Name | Symbol | Entrezl | Defined_Genelists | Filter | |
| 20 | hum_alu_at | | | | | TRUE | |
| 39 | AFFX-HUM | signal trar | STAT1 | 6772 | Apoptotic Signaling in Respo | TRUE | |
| 43 | AFFX-HUM | glyceralde | GAPDH | 2597 | Downregulated of MTA-3 in E | TRUE | |
| 44 | AFFX-HUM | glyceralde | GAPDH | 2597 | Downregulated of MTA-3 in E | TRUE | |
| 45 | AFFX-HUM | glyceralde | GAPDH | 2597 | Downregulated of MTA-3 in E | TRUE | |
| 46 | AFFX-HSA | actin, beta | ACTB | 60 | Chromatin Remodeling by hS | TRUE | |
| 47 | AFFX-HSA | actin, beta | ACTB | 60 | Chromatin Remodeling by hS | TRUE | |
| 48 | AFFX-HSA | actin, beta | ACTB | 60 | Chromatin Remodeling by hS | TRUE | |
| 52 | AFFX-M27830_5_at | | | | | TRUE | |
| 55 | AFFX-HSA | actin, beta | ACTB | 60 | Chromatin Remodeling by hS | TRUE | |
| 58 | AFFX-HUM | glyceralde | GAPDH | 2597 | Downregulated of MTA-3 in E | TRUE | |
| 64 | AB000220_ | sema dom | SEMA3C | 10512 | Axon guidance | TRUE | |
| 68 | AB000460_ | family wit | FAM193A | 8603 | | TRUE | |
| 70 | AB000464_ | NOP14 an | NOP14-AS | 317648 | | TRUE | |
| 73 | AB000468_ | ring finger | RNF4 | 6047 | | TRUE | |
| 79 | AB001106_ | glia matur | GMFB | 2764 | | TRUE | |
| 80 | AB001325_ | aquaporin | AQP3 | 360 | | TRUE | |
| 82 | AB002315_ | KIAA0317 | KIAA0317 | 9870 | | TRUE | |
| 86 | AB002380_ | Rho guani | ARHGEF12 | 23365 | Axon guidance, Regulation o | TRUE | |
| 87 | AB002382_at | | | | | TRUE | |
| 90 | AB003102_ | proteason | PSMD11 | 5717 | Proteasome | TRUE | |
| 91 | AB003103_ | proteason | PSMD12 | 5718 | Proteasome | TRUE | |
| 92 | AB003177_ | proteason | PSMD9 | 5715 | | TRUE | |
| 93 | AB003698_ | cell divisi | CDC7 | 8317 | Cell cycle | TRUE | |
| 94 | AB004884_ | tousled-li | TLK2 | 11011 | | TRUE | |
| 99 | AC000064_ | GATA zinc | GATAD1 | 57798 | | TRUE | |

# The collated project workbook

## Filtered log ratio or log intensity sheet

View the matrix of log-expression data with data filters applied.

B2 3.573403335845947

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Display the data | Brain_MD_1 | Brain_MD_2 | Brain_MD_3 | Brain_MD_4 | Brain_MD_5 | Missing | P-Value | Rank | Variance | Num 1.5-Fold | Absent (Affy) | Filter | |
| 1 | ProbeSet | | | | | | | | | | | | | |
| 20 | hum_alu_at | 14.27817 | 14.44255 | 15.29377 | 14.9108 | 16.8653 | 0 | | | | 57 | 3 | TRUE | |
| 39 | AFFX-HUMISGF3A/M97 | 8.927432 | 8.714246 | 8.096399 | 7.865125 | 10.75989 | 0 | | | | 53 | 22 | TRUE | |
| 43 | AFFX-HUMGAPDH/M33 | 14.27731 | 10.03617 | 14.0167 | 13.4024 | 13.36646 | 0 | | | | 18 | 0 | TRUE | |
| 44 | AFFX-HUMGAPDH/M33 | 14.1479 | 10.66977 | 14.57995 | 13.45807 | 13.64926 | 0 | | | | 26 | 1 | TRUE | |
| 45 | AFFX-HUMGAPDH/M33 | 14.28507 | 13.62297 | 15.00179 | 14.45779 | 14.12614 | 0 | | | | 28 | 0 | TRUE | |
| 46 | AFFX-HSAC07/X00351_5 | 14.13774 | 9.240791 | 14.31056 | 13.54756 | 14.78994 | 0 | | | | 34 | 2 | TRUE | |
| 47 | AFFX-HSAC07/X00351_M | 14.61301 | 11.43411 | 14.84308 | 14.13051 | 15.13767 | 0 | | | | 36 | 2 | TRUE | |
| 48 | AFFX-HSAC07/X00351_3 | 14.08011 | 14.59584 | 14.15988 | 13.69802 | 14.62119 | 0 | | | | 26 | 0 | TRUE | |
| 52 | AFFX-M27830_5_at | 10.9794 | 9.914385 | 11.43107 | 8.792543 | 11.89633 | 0 | | | | 78 | 38 | TRUE | |
| 55 | AFFX-HSAC07/X00351_3 | 8.794507 | 10.46557 | 10.64404 | 9.315951 | 11.88684 | 0 | | | | 54 | 5 | TRUE | |
| 58 | AFFX-HUMGAPDH/M33 | 7.084365 | 9.240791 | 8.576831 | 8.371586 | 9.057992 | 0 | | | | 66 | 44 | TRUE | |
| 64 | AB000220_at | 9.557537 | 7.882643 | 6.681361 | 7.042521 | 9.971543 | 0 | | | | 61 | 34 | TRUE | |
| 68 | AB000460_at | 10.15234 | 10.83368 | 10.6047 | 11.35548 | 10.23002 | 0 | | | | 32 | 1 | TRUE | |
| 70 | AB000464_at | 8.536878 | 10.04576 | 8.547344 | 9.687808 | 10.29692 | 0 | | | | 53 | 27 | TRUE | |
| 73 | AB000468_at | 9.438828 | 7.761551 | 8.468632 | 3.445586 | 9.108524 | 0 | | | | 52 | 15 | TRUE | |
| 79 | AB001106_at | 7.702686 | 9.647458 | 10.19205 | 6.994023 | 8.816983 | 0 | | | | 55 | 14 | TRUE | |
| 80 | AB001325_at | 8.140219 | 10.37069 | 9.846106 | 7.858368 | 10.03342 | 0 | | | | 37 | 12 | TRUE | |
| 82 | AB002315_at | 6.790634 | 7.066089 | 6.842642 | 3.445586 | 7.569856 | 0 | | | | 54 | 36 | TRUE | |
| 86 | AB002380_at | 8.591326 | 8.169925 | 6.870123 | 6.662817 | 9.562243 | 0 | | | | 68 | 23 | TRUE | |
| 87 | AB002382_at | 10.56322 | 9.569856 | 8.764255 | 8.778295 | 9.473705 | 0 | | | | 39 | 32 | TRUE | |
| 90 | AB003102_at | 10.52527 | 9.098032 | 9.39959 | 7.803138 | 10.15861 | 0 | | | | 54 | 7 | TRUE | |
| 91 | AB003103_at | 6.743329 | 7.721099 | 8.689096 | 6.994023 | 8.523562 | 0 | | | | 54 | 18 | TRUE | |
| 92 | AB003177_at | 8.568888 | 8.430452 | 7.814628 | 7.32333 | 9.092757 | 0 | | | | 45 | 26 | TRUE | |
| 93 | AB003698_at | 6.821331 | 8.144658 | 7.778827 | 5.515975 | 8.754888 | 0 | | | | 51 | 39 | TRUE | |

# The collated project workbook

## Gene annotations worksheet (Optional)

Contains gene annotations which were automatically downloaded from the Affymetrix or SOURCE database using the annotations tool.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ProbeSet (Double-click) | Name | Accessi | UGClus | Symbol | EntrezI | Chrom | Cytoba | GO | Filter | |
| 20 | hum_alu_at | | U14573 | | | | | | | TRUE | |
| 39 | AFFX-HUN | signal trar | M97935 | Hs.642990 | STAT1 | 6772 | 2 | 2q32.2 | Biological | TRUE | |
| 43 | AFFX-HUN | glyceralde | M33197 | Hs.544577 | GAPDH | 2597 | 12 | 12p13 | Biological | TRUE | |
| 44 | AFFX-HUN | glyceralde | M33197 | Hs.544577 | GAPDH | 2597 | 12 | 12p13 | Biological | TRUE | |
| 45 | AFFX-HUN | glyceralde | M33197 | Hs.544577 | GAPDH | 2597 | 12 | 12p13 | Biological | TRUE | |
| 46 | AFFX-HSA | actin, beta | X00351 | Hs.520640 | ACTB | 60 | 7 | 7p22 | Molecular | TRUE | |
| 47 | AFFX-HSA | actin, beta | X00351 | Hs.520640 | ACTB | 60 | 7 | 7p22 | Molecular | TRUE | |
| 48 | AFFX-HSA | actin, beta | X00351 | Hs.520640 | ACTB | 60 | 7 | 7p22 | Molecular | TRUE | |
| 52 | AFFX-M27830_5_at | | M27830 | | | | | | | TRUE | |
| 55 | AFFX-HSA | actin, beta | X00351 | Hs.520640 | ACTB | 60 | 7 | 7p22 | Molecular | TRUE | |
| 58 | AFFX-HUN | glyceralde | M33197 | Hs.544577 | GAPDH | 2597 | 12 | 12p13 | Biological | TRUE | |
| 64 | AB000220 | sema dom | AB000220 | Hs.269109 | SEMA3C | 10512 | 7 | 7q21-q31 | Biological | TRUE | |
| 68 | AB000460 | family wit | AB000460 | Hs.652364 | FAM193A | 8603 | 4 | 4p16.3 | Cellular C | TRUE | |
| 70 | AB000464 | NOP14 an | AB000464 | Hs.398178 | NOP14-AS | 317648 | 4 | 4p16.3 | | TRUE | |
| 73 | AB000468 | ring finger | AB000468 | Hs.740360 | RNF4 | 6047 | 4 | 4p16.3 | Molecular | TRUE | |
| 79 | AB001106 | glia matur | AB001106 | Hs.151413 | GMFB | 2764 | 14 | 14q22.2 | Molecular | TRUE | |
| 80 | AB001325 | aquaporin | AB001325 | Hs.234642 | AQP3 | 360 | 9 | 9p13 | Biological | TRUE | |
| 82 | AB002315 | KIAA0317 | AB002315 | Hs.730659 | KIAA0317 | 9870 | 14 | 14q24.3 | Molecular | TRUE | |
| 86 | AB002380 | Rho guani | AB002380 | Hs.24598 | ARHGEF12 | 23365 | 11 | 11q23.3 | Molecular | TRUE | |
| 87 | AB002382_at | | AB002382 | | | | | | Biological | TRUE | |
| 90 | AB003102 | proteason | AB003102 | Hs.595584 | PSMD11 | 5717 | 17 | 17q11.2 | Biological | TRUE | |
| 91 | AB003103 | proteason | AB003103 | Hs.592689 | PSMD12 | 5718 | 17 | 17q24.2 | Biological | TRUE | |
| 92 | AB003177 | proteason | AB003177 | Hs.131151 | PSMD9 | 5715 | 12 | 12q24.31- | Biological | TRUE | |
| 93 | AB003698 | cell divisi | AB003698 | Hs.533573 | CDC7 | 8317 | 1 | 1p22 | Biological | TRUE | |
| 94 | AB004884 | tousled-li | AB004884 | Hs.445078 | TLK2 | 11011 | 17 | 17q23 | Biological | TRUE | |

# Part IV:

# Data filtering and normalization options

# [Hands-on instructions]

[Data filtering-Pomeroy]

1. Click on **ArrayTools → Re-Filter, normalize and subset the data**.
2. Click on the four buttons **Spot filter**, **Normalization**, **Gene filter and Gene Subset** at the TOP of the form, to see the available options and view the current settings applied on the dataset.
3. By clicking “OK” the default filtering and normalization is performed on the data set.

# Data filtering options

## Single-Channel: Spot filter

- Intensity filter:  May filter out spots with low intensity in single channel or threshold low intensity in forming log intensities.
- Detection Call: Exclude a probeset if the Detection call value is “A”,”M” ,“P” or “No Call”.
- Dual channel: Background correction and averaging replicate spots can be performed.

# Data filtering options

## Normalization and truncation

- Normalization and truncation steps are applied *after* data has been spot-filtered, but *before* screening out genes
- Arrays are normalized before outlying expression levels are truncated.
- Purpose of truncation is primarily to prevent extremely large ratios from being formed by small denominators in dual-channel data. The truncation option is useful if the dual-channel intensities have not been thresholded.

# Data filtering options

## Data transformation options

- Normalization:

  For single-channel data: Default option is to median-center all arrays to a reference array, based on all genes or only a set of housekeeping genes. The reference array may be explicitly chosen, or a “median” array can be automatically found.

- Truncation: Truncate extreme values (large log-intensities for single-channel data, or large absolute log-ratios for dual channel data)

# Data filtering options

## Gene filters: Gene variation

- Fold-change filter: Specify a minimum percentage of log-expression values which must meet a specified fold-change criteria
- Log-ratio (or log-intensity) variation filter:

  Screen genes which do not vary much over the set of samples:

  1. Significance criterion compares the variance of each gene against the "average" gene
  2. Percentile criterion screens a specified percentage of genes with smallest variance

# Data filtering options

## Gene filters: Gene quality

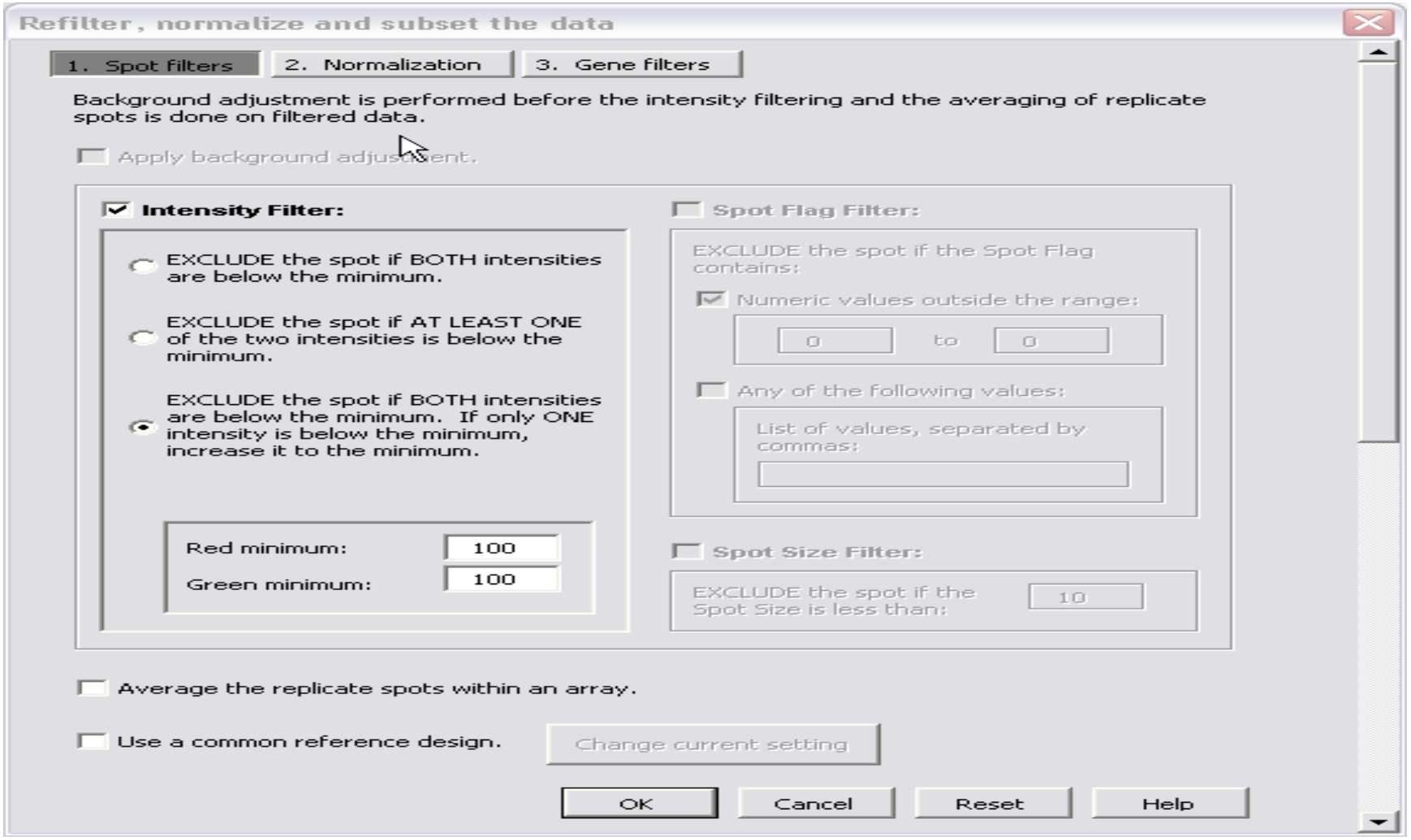
- Missing value filter: Screens out genes which contain too many missing values over the set of samples
- Percent absent filter: For Affymetrix data, can filter out a probeset if too many expression values had an Absent call
- Minimum Intensity: This option is only available for single channel data. It filters out genes whose $50^{th}$ percentile normalized log intensity is less than the log of the user defined value.

# Data filtering options

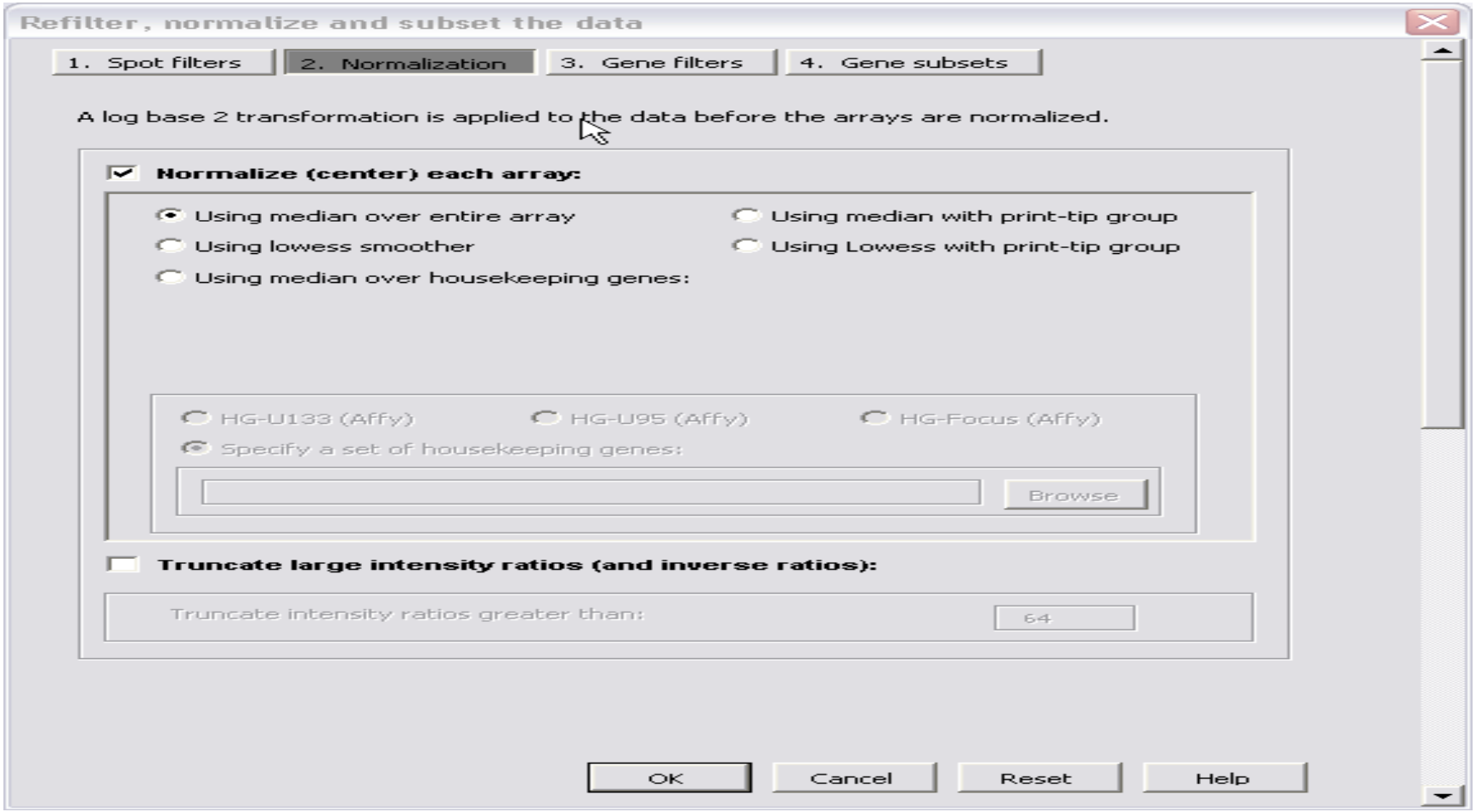
## Gene subsets

- Select genelists for analysis: User may subset the data by selecting one or more genelists to INCLUDE or EXCLUDE. If more than one genelist is selected, then the UNION of all genes on those genelists will be used.
- Specify gene labels to exclude: User may exclude genes based on gene identifier labels. For example, all genes with “Empty” in the gene description field may be excluded.
- CAUTION: Gene subsetting is applied globally to the entire dataset, not just to a specific analysis.
- Probe reduction: Reduce multiple probe sets per gene by choosing the most variably expressed or the maximally expressed probe/probeset.

# Spot Filtering options- Dual Channel data



# Normalization options- Dual channel data



# Part V:

# Overview of some analysis tools

# Scatterplot tools

- Scatterplot of experiment v. experiment: Plots intensity, geometric mean of the red and green intensities, and intensity ratio on log-scale. The M-A plot can be implemented for two-channel data as a plot of the log-ratio versus the average log-intensity.
- Scatterplot of phenotype averages: Plots averages over experiment classes
- Online demo http://linus.nci.nih.gov/PowerPointSlides/Scatterplot.wmv

# [Optional: Hands-on instructions]

## [Scatterplot of phenotype averages]

1. Now click on **ArrayTools → Graphics -> Scatterplot → Phenotype averages**.
2. Select the variable **Dx** as the phenotype class to average over, and then click **OK**.
3. This launches a 2-D and 3-D scatter plot.
4. Right click on the 2-D plot to modify scatter plot properties, select up/down regulated genes as well as link genes in other plots.

# [Optional: Hands-on instructions]

[Scatterplot of experiment v. experiment-Pomeroy Data]

1. Click on **ArrayTools → Graphics ->Scatterplot → Array vs. Array.**
2. Select **Log(Intensity)** for the **Brain_MD_1** experiment for the X-values and **Log(Intensity)** for the **Brain_MD_MGlio_1** experiment as Y-values.
3. Select “2” as the number of panels.
4. Click “OK”. Then, right click on the plot to change scatterplot properties, select up/down regulated genes etc.

# Analysis Wizard

- Click on “ArrayTools” pull down menu.
- Select “Analysis Wizard”
- Our research interest is to find genes that are differentially expressed among pre-defined classes of samples.

# Analysis Wizard

- **Finding Genes**

Finding differentially expressed genes/gene sets amongst classes.

- **Prediction**

Develop a classifier for predicting the class of a sample

- **Clustering/Visualizing**

Visualizing/Clustering of Genes and Samples.

# Finding Genes

- Comparing classes (Class Comparison)
- Correlated with a quantitative trait (Quantitative Trait Analysis)
- Correlated with survival (Survival Analysis)
- Time Course Analysis (Plug-in)

# Tools for finding Genes/Genesets comparing classes

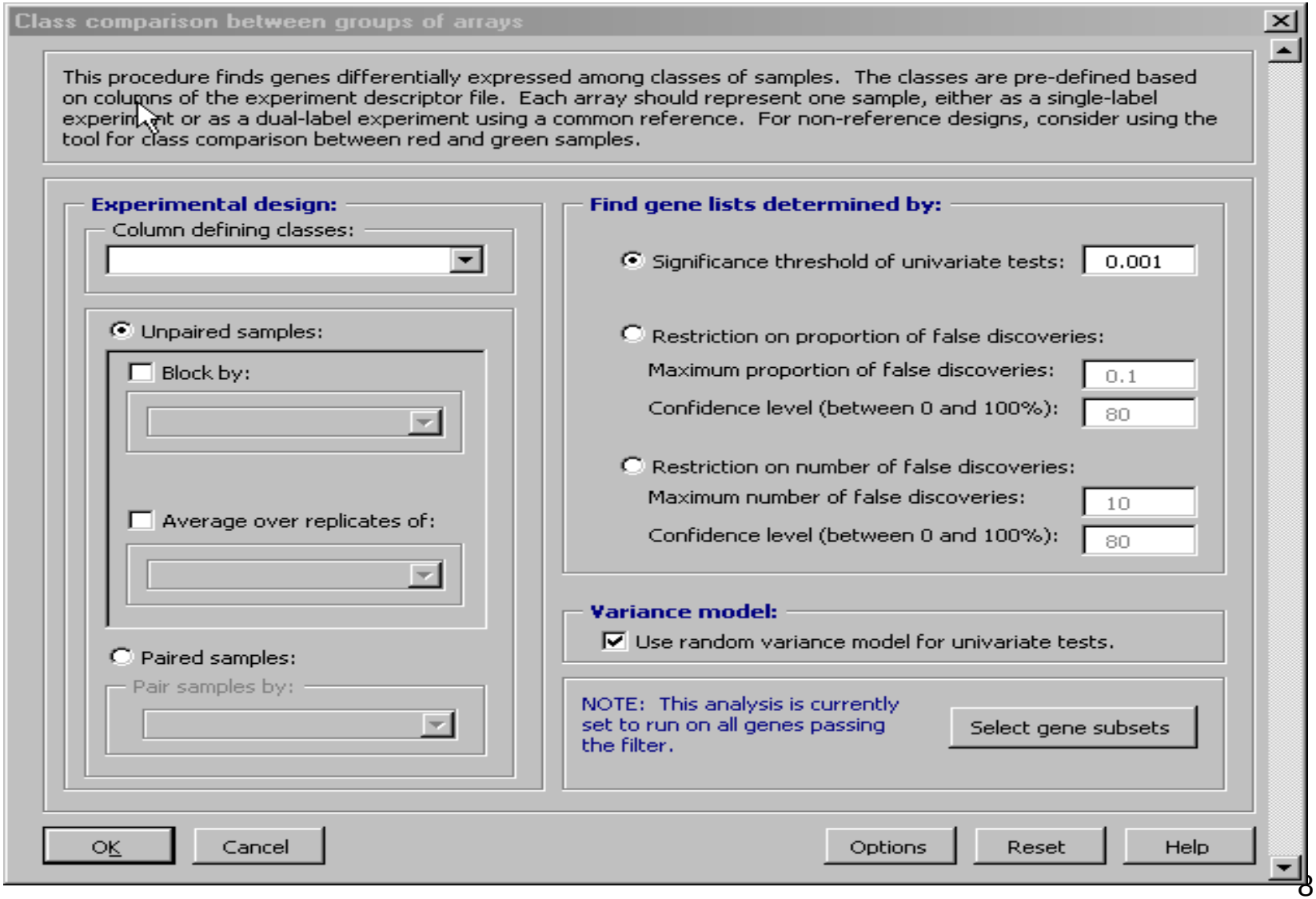
- Class Comparison Between groups of arrays
- SAM
- Gene Set Expression Comparison.
- ANOVA models

# Class comparison tool

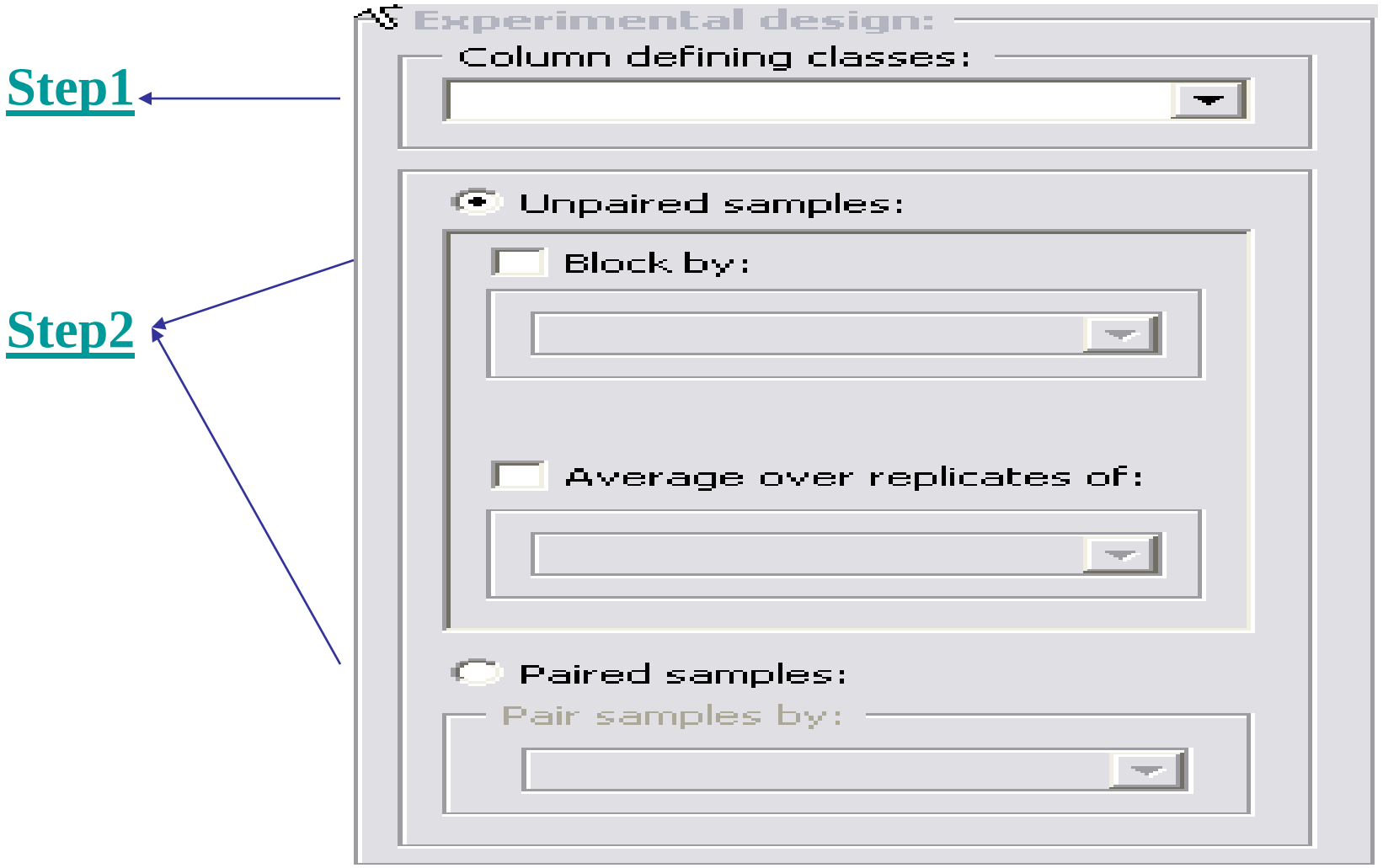
Between groups of arrays

- FOR SINGLE-CHANNEL DATA, OR DUAL-CHANNEL REFERENCE DESIGNS.
- Class comparison tool uses univariate t/F-tests, with multivariate permutation tests
- Permutation tests are nonparametric, and take correlation among genes into account
- Paired analysis option
- Produces a gene list which can be used for further analysis.
- Produces chromosomal distribution and GO analysis if genes have already been annotated using the Affymetrix or SOURCE database.

# Class Comparison



# Class Comparison
# Experimental design



# Class comparison tool

1. Enter the class column from the ‘Experiment descriptor’ worksheet that defines the classes for the samples.
2. Specify if this is a paired or un-paired analysis. An analysis is said to be paired if for example, you have the same sample from a patient before and after a treatment. You then need a column in the experiment descriptor worksheet that will contain identical values for pair of arrays.
3. If this is an unpaired analysis, do you have a blocking factor?
4. If this is an unpaired analysis, do you have an replicates you want to average across?

# Class comparison tool
# Blocking Factor

Experimental designs containing a blocking factor can be performed by specifying which column in the Experiment descriptor worksheet contains a blocking variable. When selected, the influence of the blocking variable is taken into consideration when analyzing the differences between classes.

Examples of variables that may be considered as Blocking factors:

- Clinical Site for patient data
- Print set for cDNA spotted arrays
- Batch of arrays

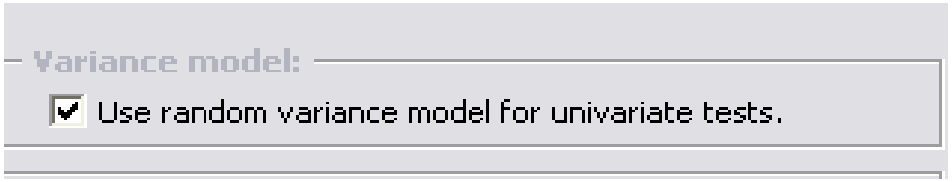
# Average over replicates

- If multiple arrays have been performed using the same sample RNA then an average of these replicates should be used instead of the individual arrays in the analysis.
- In the ‘experiment descriptor’ worksheet, there should be column containing sample ids for these arrays.
- Arrays that contain the identical values of the sample id variable are considered as replicates and will be averaged in the analysis.

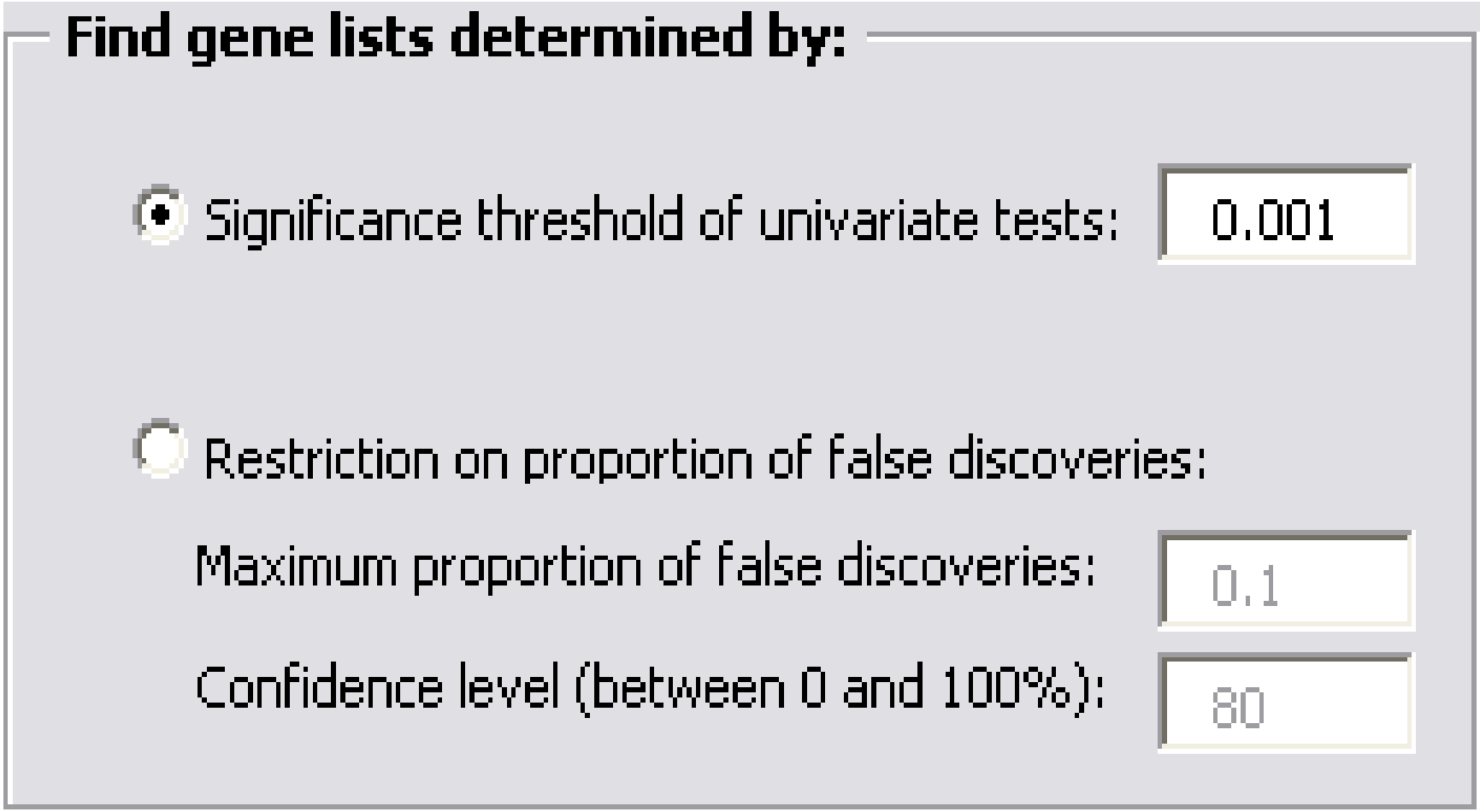
# Class comparison tool

## Random variance option

- The random variance test has more power because the “average” variance in the denominator adds degrees of freedom for the test statistic.
- Should be used for small sample sizes.
- Dialog option:



# Find genes lists determined by:



# Class comparison tool

## Univariate significance test

- Compute the univariate p-value for each gene, and sort list of genes by smallest p-value.
- In the univariate setting (i.e., testing significance of one gene at a time), the p-value is defined to be the probability of obtaining a false positive result.
- However, once a list of univariately significant genes is found, it is not clear how many of those genes are false positives.

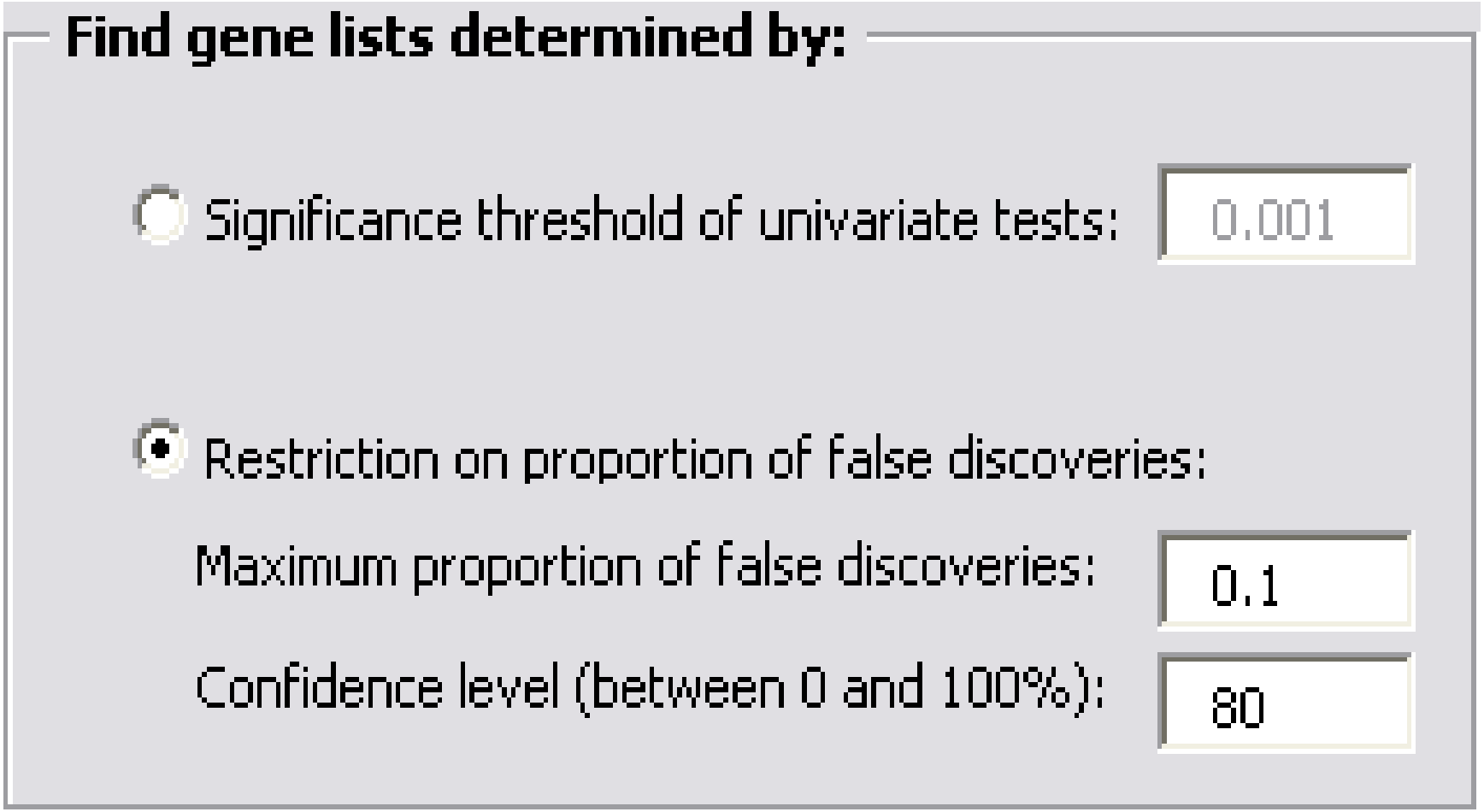
# [Hands-on instructions]

## [Class comparison – univariate significance threshold]

1. Using the Pomeroy data, run the Class Comparison tool by clicking on **ArrayTools → Class comparison → Between groups of arrays**.
2. Select the **Medulo vs Glio** variable as the column defining the classes. Select the **Random variance model** option, and select the **Significance threshold of univariate tests: 0.001**.
3. Leave all other options at default levels. Now click **OK** on the main dialog to launch the analysis.
4. You will see a DOS window appear in your Windows Task Bar at the bottom of your screen. If you click on the DOS window, you can monitor the analysis running inside the DOS window.
5. When the analysis has completed, it will automatically open up an HTML file which displays the output.

# Class comparison tool

## Multivariate permutation test



# Class comparison tool

## Multivariate permutation test

- In the multivariate setting (i.e., when testing many genes for significance at the same time), ask the question:  What p-value cutoff should I use to guarantee that 90% of the time, I get less than P proportion of false positives (where P is specified by the user)?
- To answer this question, we compute the permutation distribution of the p-value cutoffs for which we would get P proportion of false positives.
- The output tells us how far down the list we would be able to go in order to be assured (with a certain confidence) of getting less than P proportion of false positives.

# [Hands-on instructions]

## [Class comparison – Restricting proportion of false positives]

1. Using the Pomeroy data, run the Class Comparison tool by clicking on **ArrayTools → Class comparison → Between groups of arrays**.
2. Select the **Medulo vs Glio** variable as the column defining the classes. Select the **Random variance model** option, and select the **Restriction on proportion of false discoveries** with **maximum proportion** = 0.1 and 90% **Confidence level**.
3. Click on the **options** and change the name of the **output** folder to “ClassComparisonMPT”
4. Leave all other options at default levels. Now click **OK** on the main dialog to launch the analysis.
5. When the analysis has completed, it will automatically open up an HTML file which displays the output.

# Gene ontology analysis

- In the class comparison, class prediction, survival analysis, or quantitative traits analysis output, the observed vs. expected frequency is computed for each Gene Ontology class represented in the selected genelist, as well as for each upstream Gene Ontology class. By default, results are printed only for classes represented by at least five genes in the selected genelist, and with an observed versus expected ratio of at least 2.

# Class comparison

## Significance Analysis of Microarrays (SAM)

- SAM is another popular method for false discovery control, which controls the ***average*** proportion of false discoveries rather than the ***probability*** of a given number or proportion of false discoveries.
- It is a slightly less stringent control than the multivariate permutation test for controlling false discoveries used in the other class comparison tools, but is included in BRB-ArrayTools because of its popularity.

# [Hands-on instructions]

## [Significance Analysis of Microarrays – Pomeroy data]

1. Still using the Pomeroy data, run the SAM tool by clicking on **ArrayTools → Class comparison → Significance Analysis of Microarrays (SAM)**.
2. Again, select the **Medulo vs Glio** variable as the column defining the classes, select the **90$^{th}$ percentile** option, and leave all other parameters at default levels.
3. Check the option to perform **Gene ontology Observed vs Expected analysis.**
4. Now click **OK** to exit the options dialog, and click **OK** on the main dialog to launch the analysis.

# Gene set Expression Comparison

- Allows users to find significant ***sets*** of genes rather than just significant genes.
- For the **Gene Ontology comparison**, all Gene Ontology classes that are represented in the data are tested for significance.
- For **Pathway Comparison**, all the pathways that are represented in the data are tested. For Human, the BioCarta or KEGG pathways are tested and for mouse, the BioCarta pathways are compared. Additionally, Broad/MIT pathways can be downloaded to be used in analyses.
- For the **User Gene Lists comparison**, the user can select specific genelists that the user would like to test for significance.
- Transcription factor target gene lists and microRNA target genelists have been added to the Gene List comparison tool.

# Gene Set Expression Comparison

- Compute p-value of differential expression for each gene in the gene set(k=number of genes)
- Compute a summary (S) of these p-values
- Determine whether the summary test (S) is more extreme than would be expected from a random sample of “k” genes on that platform.
- Two types of summaries provided:
  - Average of log p-values
  - Kolmogrov-Smirnov statistic.

# Efron-Tibshirani's GSA maxmean test

- Tests the null hypothesis that for a gene set the average degree of differential expression is greater than expected from a random set of genes.
- Uses the maxmean statistic as follows:
- Take the di scores for all the genes within a geneset.
- Set negative scores to 0 and compute ‘avpos’ as the average of the positive scores and zeros.
- Similarly set the positive scores to 0 and compute the ‘avneg’ as the averages of the negative scores and zeros.
- A gene set is scored ‘avpos’ if |avpos| > |avneg| or else the gene set is scored ‘avneg’

# [Hands-on instructions]

## [Class Comparison – Pathway Comparison: Pomeroy data]

1. On the Pomeroy data, run the Class Comparison tool by clicking on **ArrayTools → Class comparison → Gene set Expression Comparison.**

2. Select the **Medulo vs Glio** variable as the column defining the classes. Select the **Random variance model** option and **Pathways**, and leave all other options at default levels. Now click **OK** on the main dialog to launch the analysis.

3. You will see a DOS window appear in your Windows Task Bar at the bottom of your screen. If you click on the DOS window, you can monitor the analysis running inside the DOS window.

4. When the analysis has completed, it will automatically open up an HTML file which displays the output.

# Quantitative trait tool

- Selects genes which are univariately correlated with a quantitative trait such as age or time point.
- Controls number and proportion of false discoveries in entire list:  uses a multivariate permutation test which takes advantage of the correlation among genes.
- Produces a gene list which can be used for further analysis.
- Produces chromosomal distribution and GO analysis if genes have already been annotated using the SOURCE database.

# Survival analysis tools

- Find Genes Correlated with Survival tool, selects genes which are univariately correlated with survival
- Controls number and proportion of false discoveries in entire list:  uses a multivariate permutation test which takes advantage of the correlation among genes
- Produces a gene list which can be used for further analysis.
- Produces chromosomal distribution and GO analysis if genes have already been annotated using the SOURCE database.

# Survival Gene Set analysis

- This analysis tool finds sets of genes for which the expression levels are correlated to survival. Similar to the Gene Set Expression comparison tool, this tool can be used to analyze Gene Ontology categories, Pathways, micro RNA targets, transcription factor targets and user defined gene lists.
- The permutation p-values from the LS and KS statistics are computed.
- The HTML output lists the sets of genes and the associated p-values.

# Classification of samples

- Cluster analysis vs. classification
- Use cluster analysis to discover new classes, or for visualization purposes
- Use classification when classes are already specified
- Classification is supervised learning, and generally has more power because it uses the known information about the hybridized samples.
- Use the Class Prediction tool when the primary interest is to form a classifier to predict the class of new samples.

# Hierarchical clustering tools

- Clustering of genes and samples produces visual image plot of log-expression data, where ordering is determined by ordering of dendrogram
- Can compute measures to assess cluster reproducibility when clustering samples alone
- May cluster based on gene subsets rather than on the entire gene set
- Interface to Cluster 3.0 and TreeView originally produced by the Stanford group is also included, and allows for easy exportation of results.

# [Hands-on instructions]

## [Cluster analysis – Pomeroy data]

1. Using the Pomeroy data set.
2. Run the cluster analysis by clicking on **ArrayTools → Clustering → Genes (and samples)**.
3. Click on the **Select gene subsets** button, and under **Select genes for analysis**, choose the **ClassComparison** genelist, and click **OK**.
4. Now click on the **Options** button, and choose **Medulo vs Glio** as the variable under **Label the experiments**. Click **OK** to exit the options dialog, and click **OK** on the main dialog to launch the analysis.

# [Hands-on instructions – cont'd]

[Cluster analysis – Pomeroy data]

5. The analysis will open up a **Cluster viewer** worksheet inside your project workbook. The first plot presented is the Heat Map image in a draft form. Using **Zoom and Recolor** button you can change the color scheme of the map. Click the button and on the dialog page select **Red/Blue** scheme and de-select the **Use quantile data…** This coloring option should look familiar to the dChip users.
6. The setting for using the quantile data ranges when distributing colors on the scale leads to the heat map when two different major colors on the map represent not the range of values of equal length but rather the sets with the equal number of points.

# [Hands-on instructions – cont'd]

[Cluster analysis – Pomeroy data]

7. You can also use **Zoom and Recolor** option to zoom in which will present the fragment of the map in a separate window and zoom out when you have too many genes for the regular map to fit into window but want to see the whole picture. Select genes 50 to 60 and arrays 6 to 30 to zoom in.
8. Right click on the one of the gene **Info** links in the left part of the IE window and select "Open in New Window"

# [Hands-on instructions – cont'd]

[Cluster analysis – Pomeroy data]

9: Use **Previous** button on ClusterViewer to get to the dendrogram plot where you can **cut the tree (# 4 clusters)**. Then you can click the **Next** button to scroll through the output plots. You can also click on **List genes** to identify the genes within each cluster. Note that the samples are ordered by default according to a hierarchical clustering of the samples. However, the dendrogram for the hierarchical clustering of the samples is not shown. To view the dendrogram for the hierarchical clustering of samples, you must run it as a separate analysis.

# [Hands-on instructions – cont'd]

[Cluster analysis – Pomeroy data]

10. Still with the Pomeroy data in front of you, click on the **ArrayTools→ Clustering → Sample alone** menu item.
11. Select the **Compute the cluster reproducibility** option
12. Now click on the **Options** button, and choose **Dx** as the variable under **Label the experiments**.
13. Click **OK** to exit the options dialog, and click **OK** on the main dialog to launch the analysis.

# [Hands-on instructions – cont'd]

[Cluster analysis – Pomeroy data]

14: The analysis will create a dendrogram plot of the hierarchical clustering of samples inside the **Cluster viewer** worksheet. You may then click the **Cut tree(# of cluster 3)** button to "cut the tree", thereby defining clusters of samples from the dendrogram. After you have defined clusters of samples by "cutting the tree", the analysis will be run in a DOS window which appears in your Windows Task Bar, and an HTML file containing the output will open up automatically once the computation is completed

# Cluster reproducibility

- Add perturbation noise to original data
- Re-cluster perturbed data to assess stability of original clusters
- Overall and cluster-specific measures
- Robustness (R) index measures the proportion of pairs of specimens within a cluster for which the members of the pair remain together in the re-clustered perturbed data
- Discrepancy (D) index measures the number of discrepancies (additions or omissions) comparing an original cluster to a best-matching cluster in the re-clustered perturbed data.

# Multidimensional scaling

- Rotating scatterplot: Gives three-dimensional visualization of relationships between samples
- Global test of clustering in samples: Compares spatial distribution of data to white noise. Large deviation from Gaussian normal distribution indicates presence of clustering.

# [Hands-on instructions]

## [Multidimensional scaling –Pomeroy data]

1. Still using the **Pomeroy** dataset, run the multidimensional scaling by clicking on **ArrayTools → Graphics -> Multidimensional scaling → of samples**.
2. Now choose **Dx** as the variable to **Color the rotating scatterplot**. click **OK** on the main dialog to launch the analysis.
3. A Java window will be launched, containing a scatterplot which can be rotated using arrow control buttons. Each point represents a sample, and points can be identified by brushing over them with your mouse.
4. A PowerPoint slide is automatically created, so that you can also launch the rotating scatterplot at a later point from PowerPoint.

# Analysis Wizard- Prediction

- Class Prediction
- PAM
- Top scoring pair plug-in
- Random Forest plug-in
- Binary Tree Prediction

# Components of Class Prediction

- C1. Feature(gene) selection

  -which genes will be included in the model.

- C2. Select model type.

  -choose prediction method (DLDA,CCP etc)

  Fit the parameters for the model.

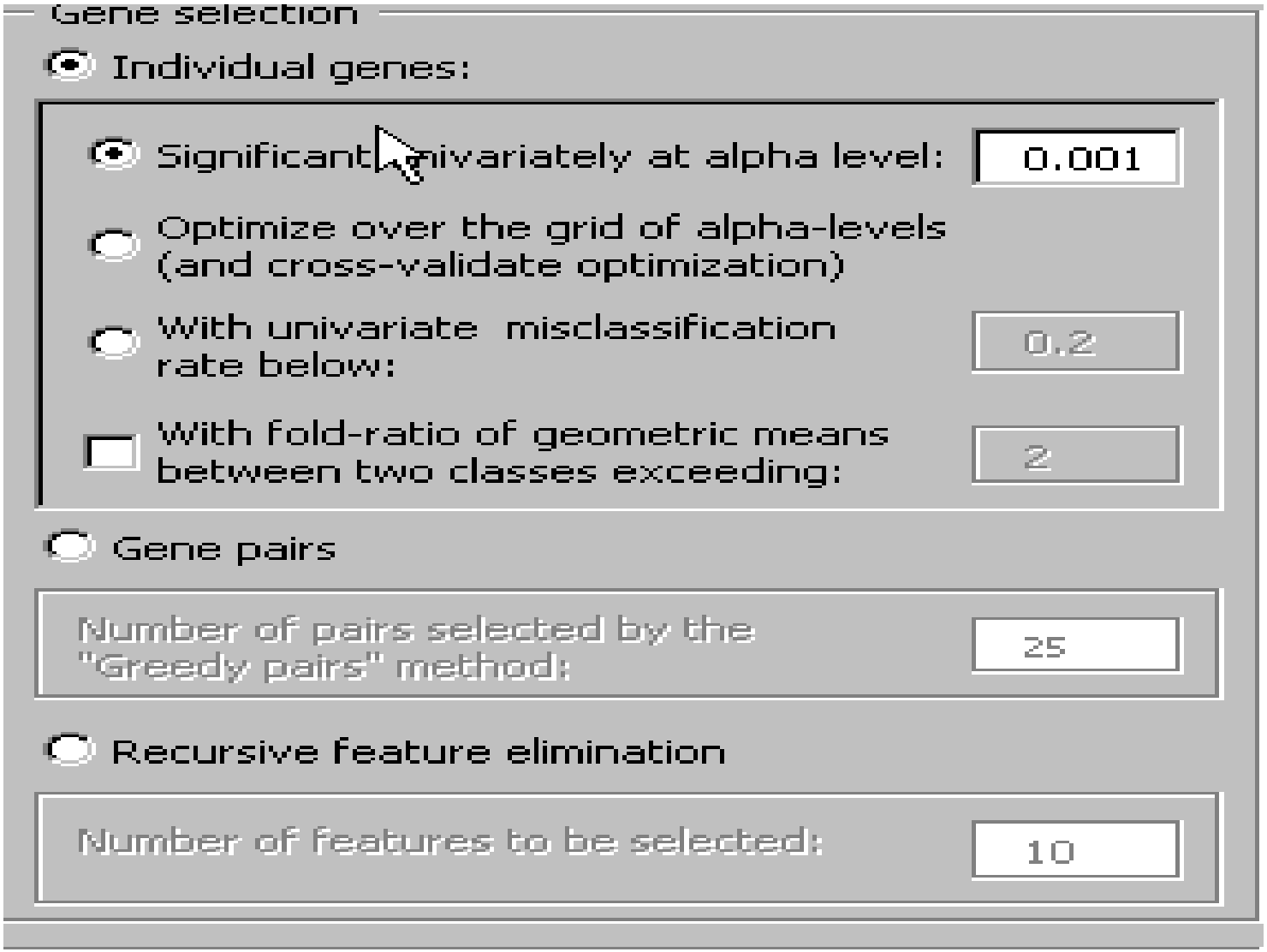
- C3. Evaluating the Classifier

  - Cross-validation

# C1. Gene Selection Criteria

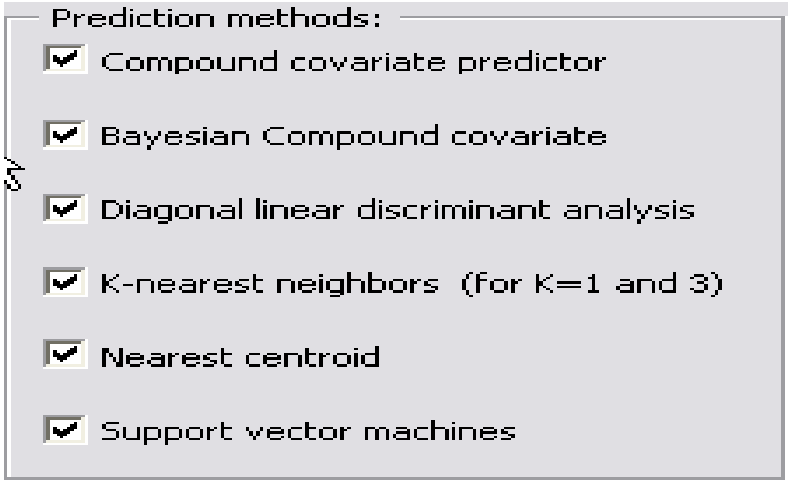
- Selection of genes may be based on univariate significance criterion or univariate misclassification rate, and minimum fold-ratio of geometric means. The univariate misclassification rate criterion is available when there are only two classes. The option to optimize over a grid of alpha values.
- In addition,we have added the option to select genes using “gene pairs”by the “greedy pair”method –Bo & Jonassen

# Gene Selection Criteria



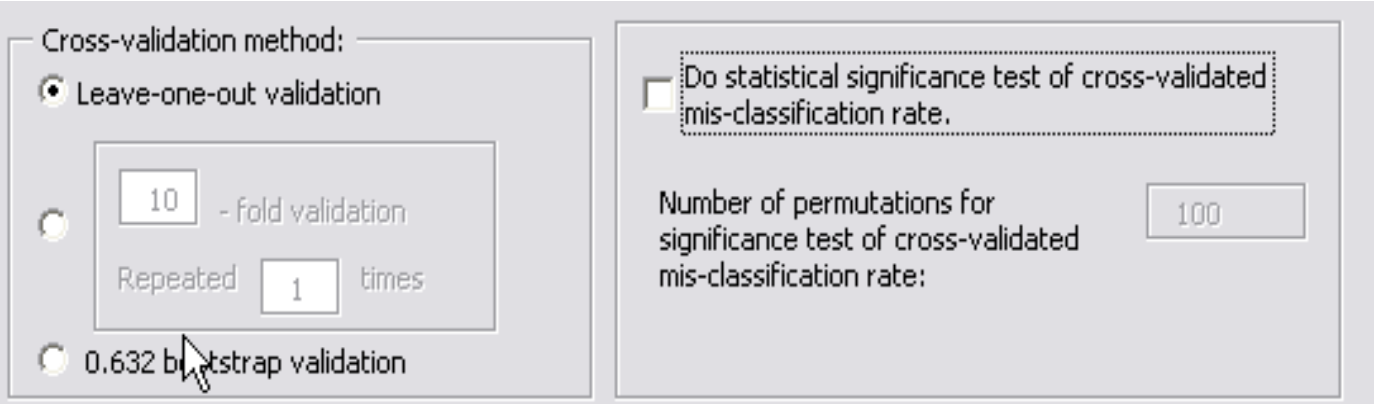
# C2. Class prediction Methods

- Six methods of prediction:
  - Compound covariate predictor (2 classes only)
  - Bayesian Compound covariate predictor (2 classes only)
  - K-nearest neighbor (2 or more classes)
  - Nearest centroid (2 or more classes)
  - Support vector machines (2 classes only)
  - Diagonal linear discriminant analysis (2 or more classes)
-



# C3. Cross-validating the classifier

- Leave-One-Out cross validation.
- K-Fold cross validation.
- +0.632 bootstrap cross-validation.
- Use leave-one-out cross-validation to compute a misclassification rate
- Re-compute the classifier, based on all but one sample
- Use the classifier to classify the sample which has been left out



# Permutation test

- Use a permutation test to assess the significance of the misclassification rate and univariate significance of each gene
- For each permutation of the class labels, re-run the cross-validation and obtain a new cross-validated misclassification rate
- The permutation p-value is based upon the rank of the misclassification rate using the original data, compared to all permutations

# Compound covariate predictor

- May only be used for classifying among two class labels
- Select genes which univariately classify the samples
- Form a compound covariate predictor as:

$$\sum_i t_i x_i \quad \left\{ \begin{array}{l} \text{where } t_i = \text{t-statistic, } x_i = \text{log-ratio,} \\ \text{and sum is taken over all significant genes} \end{array} \right.$$

- Determine the cutpoint of the predictor as the midpoint between its mean in one class and its mean in the other class

# Linear classifiers for two classes

$$l(\underline{x}) = \sum_{x \in F} w_i x_i$$

$\underline{x} =$ vector of log ratios or log signals

$F =$ features (genes) included in model

$w_i =$ weight for i - th feature

decision boundary $l(\underline{x}) >$ or $<$ cutoff

# Linear classifiers for two classes

- Diagonal linear discriminant analysis (DLDA)
- Compound covariate predictor

  - Bayesian compound covariate
- Support vector machine

# Diagonal linear discriminant analysis

- May be used for classifying among two or more class labels
- Use F-test to screen for genes which are univariately significant in classifying the samples
- Seeks a linear combination of the variables which has a maximal ratio of the separation of the class means to the within-class variance, where genes are assumed to be uncorrelated

# Bayesian Compound Covariate

- Compound Covariate score is computed for all the samples in the cross-validated training set.
- The CCP-scores of samples in each class of the training set are assumed to be from a Gaussian distribution.
- If prior probabilities are ½ - the BCCP is similar to the CCP.

# K-nearest neighbor

- May be used for classifying among two or more class labels
- Use F-test to screen for genes which are univariately significant in classifying the samples
- For k=1 and k=3, finds the k-nearest neighbors in terms of Euclidean distance over only those genes which were univariately significant
- Classify based on the majority vote of the class labels of the k-nearest neighbors

# Nearest centroid

- May be used for classifying among two or more class labels
- Use F-test to screen for genes which are univariately significant in classifying the samples
- Compute the centroid of each class as a mean over all the training samples with that class label
- Classify test sample to be same class label as the nearest centroid, using Euclidean distance over only those genes which were univariately significant

# Support vector machines

(V. Vapnik)

- Implemented only for classifying among two class labels
- Select genes which univariately classify the samples
- The SVM predictor is implemented as a linear function of the log-ratios or the log-intensities over the significant genes, that best separates the data subject to penalty costs on the number of specimens misclassified.

# Class prediction tool

## Class prediction vs. binary tree prediction

- The class prediction tool has more options:  may select all prediction methods simultaneously, may use paired samples, may use randomized variance option.
- The binary tree prediction tool splits the classes into groups of subclasses.  At each node in the tree, the binary tree prediction tool decides how to split the classes into two groups based on either a leave-one-out or a K-fold cross-validation.  The binary tree prediction tool may be useful if there is a hierarchical structure to the classes.
- However, the binary tree prediction may be very slow for a large number of samples.  Therefore, a K-fold cross-validation should be used if the number of samples is large.
- Currently the tool is limited to five classes, and requires at least four samples per class for good prediction.

# Prediction Analysis Microarray PAM

- Uses Shrunken Centroid algorithm developed by Tibshirani's group (Stanford).
- Similar to Nearest Centroid but the centroids are shrunk towards each other based on shrinking the class means for each gene towards an overall mean.
- Amount of shrinking is determined by a tuning parameter delta and the number of genes included in the classifier is determined by the value of delta.

# Important notes

- Cross validation is only valid if the test set is not used in any way in the development of the model.
- With proper CV, the model must be developed from scratch for each leave-one-out training set. This means that feature selection must be repeated for each leave-one-out.

# [Hands-on instructions ]

[Class prediction –Pomeroy data]

1. Run the Class Prediction tool by clicking on **ArrayTools → Class prediction → Class prediction**.
2. Select the **Medulo vs Glio** variable as the column defining the classes. Check the box for using the Random Variance Model.
3. Choose the univariate significance alpha=0.001.
4. Select **Options,** check the box for **Use separate test set**, and select the column “TrainingSet”.
5. Leave all other options at default levels, and click **OK**.
6. Note the Array Ids which have been misclassified by all methods.

# Plug-in utility

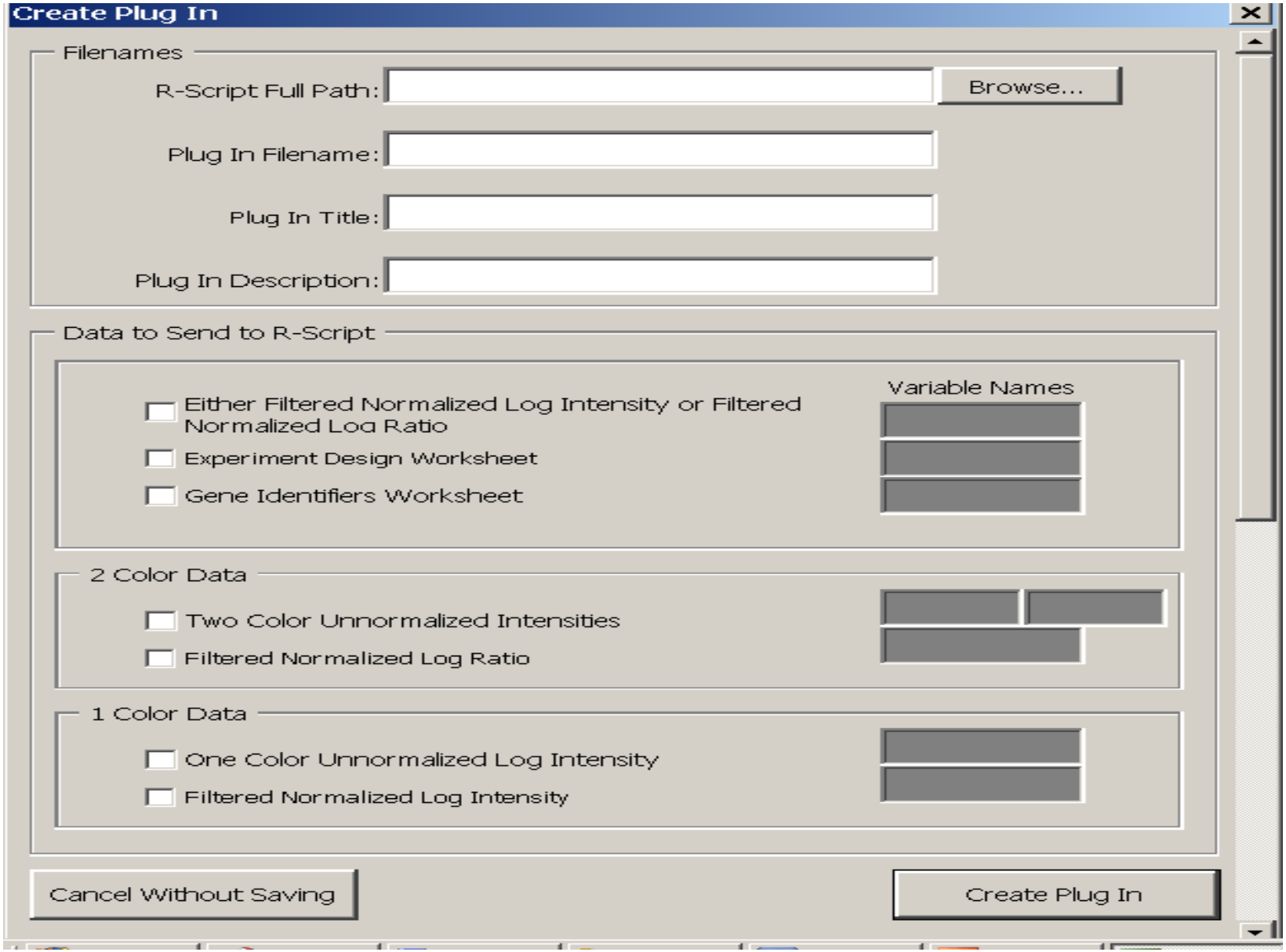
- A plug-in utility now allows users to create their own tools by writing their own scripts written in the R language
- Tools created using the plug-in utility can be distributed to other users, and added to the Plugin menu
- The user-created plug-ins are stored in the Plugins folder of the ArrayTools installation folder

# Included plugins

- Analysis of Variance – Up to four-way ANOVA. Options to include blocking factors or use random variance model.
- ANOVA of log intensities – For dual-channel non-reference designs, model includes gene-specific array effect, dye effect, and class effect. Option to use random variance model.
- ANOVA for Mixed Effects Model – Allows up to three fixed effects and one random effect.
- M vs A plot – For dual-channel data, plots log-ratio vs average log-intensity for all arrays.
- Pairwise correlation – Plots heat map showing the matrix of pairwise correlations among all arrays.
- Smoothed CDF – Plots smoothed cumulative distribution function of log-red and log-green, or log-ratio for all arrays.
- Export 1- and 2-color data to R – Exports data from Project Workbook to files which can be imported into R.

# [Additional Plugins]

- Class Prediction using TopScoring Pairs: This plugin is a different tool for class prediction by using the top-scoring pairs (TSP) classifier developed by Geman et al.
- Random Forest: This tool is another alternative to class prediction and the random forest is built from the ensemble learning method - methods that generate many classifiers and aggregate their results. The random forest is robust against overfitting and has been demonstrated to have performance competitive with the other classifiers.
- TimeSeries: This plug-in can be used for regression analysis of time series expression data.


Create Plug In
Filenames
R-Script Full Path:
Browse...
Plug In Filename:
Plug In Title:
Plug In Description:
Data to Send to R-Script
Variable Names
Either Filtered Normalized Log Intensity or Filtered Normalized Log Ratio
Experiment Design Worksheet
Gene Identifiers Worksheet
2 Color Data
Two Color Unnormalized Intensities
Filtered Normalized Log Ratio
1 Color Data
One Color Unnormalized Log Intensity
Filtered Normalized Log Intensity
Cancel Without Saving
Create Plug In

# Part VI:

# Independent practice (if time permits)

# Further help

- We hope this class has been helpful to you. This class was not designed to be comprehensive, but only an introductory overview of the features in BRB-ArrayTools. More information about the software may be obtained from the User's Manual (may be viewed by clicking on **ArrayTools -> Support -> Manuals -> User's Manual**).

- Supplementary material on analysis algorithms may be found in the BRB technical reports: http://linus.nci.nih.gov/~brb/TechReport.htm

# Acknowledgements


- Dr. Richard Simon and Biometrics Research Branch staff members.
- BRB-ArrayTools development team (past and present).
- User community!!

# Technical support

- For questions of a general nature, post a message to the BRB-ArrayTools Message Board:

  http://linus.nci.nih.gov/cgi-bin/brb/board1.cgi

- To report bugs, send email to arraytools@emmes.com

# Feedback on this class

- Please fill out a feedback form before you leave the class.
- Please make your comments specific enough to enable us to adjust this presentation for future classes.
- Thank you for participating in this class!!

# Exercise Section

**Using the breast tumors sample data set, find genes that are differentially expressed for patients before and after treatment:**

- **Obtain a gene list that contain no more than 40% of False discoveries with 95% confidence.**
- **Choosing an alternative method to the Multivariate Permutation test to control for false discoveries obtain another gene list with a 95% confidence level and controlling for 40% False discoveries.**
- **Using all genes in this sample dataset, run a scatter plot of phenotype averages with 2 fold difference and comment on the up/downward regulated genes.**