

A Tutorial on Data Analysis Using BRB-ArrayTools Version 3.5

Supriya Menezes (arraytools@emmes.com)
September 7th 2006

Class Agenda

- I. What is BRB-ArrayTools?
- II. The collated project workbook
- III. Data filtering and normalization options
- IV. Overview of some analysis tools
 - 1. Scatterplots
 - 2. Class comparison and GeneSet expression Comparison
 - 3. Clustering
 - 4. Class Prediction, Survival analysis, quantitative traits
 - 5. Multidimensional scaling
 - 6. Plugins
- V. Getting your data into BRB-ArrayTools: creating a project workbook
- VI. Questions

Part I:

What is BRB-ArrayTools?

BRB-ArrayTools

An Integrated Software Tool for DNA Microarray 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 2.3.0 or higher. 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. Compute RMA/GC-RMA probeset summaries and normalization. BRB-ArrayTools was designed to analyze a *set* of arrays rather than a single array.
- 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 plugin analysis tools within BRB-ArrayTools.

Limitations of BRB-ArrayTools

- Available only on the PC, not Macintosh.
- Analyze up to approximately 65,000 genes and an unlimited number of arrays. However, practical limitations of individual computer capacities may limit datasets to a smaller size.
- Importation of Affymetrix CEL files using RMA/GC-RMA method requires a large memory capacity even for relatively large sets of arrays and may further limit the number of arrays which can be imported.

Part II:

The collated project workbook

[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.

[Hands-on instructions]

[Importing Pomeroy Data set]

- Click on **ArrayTools** → **Import data** → **Specialformat:Affymetrix GeneChips** → **Probeset level data**

The screenshot shows a Windows-style dialog box titled "Collate Affymetric GeneChip data". It contains several sections for configuring data import:

- Introduction:** A text box stating, "This dialog form is used for collating tab-delimited text files containing probeset-level data which came from MAS 4.0 or 5.0."
- Expression data:** Two radio button options:
 - ☐ Data for each chip type is stored in a separate folder. Each folder contains a separate file for each array.
 - ☒ For each chip type, expression data for all arrays is stored in a horizontally aligned file.
- Chip Type:** Two dropdown menus:
 - "GeneChip set:" with "HuGeneFL" selected.
 - "Number of chip types used from this set:" with "1" selected.
- Probe Set ID:** Two radio button options:
 - ☒ Import annotations automatically.
 - ☐ Use my own gene identifiers file.A "File:" text box and a "Browse" button are present for the second option.
- Experiment descriptors:** A checkbox labeled "I do not have an experiment descriptor file, please create a template with just array ids." is unchecked. Below it, a "File:" text box contains the path "C:\BRB-Projects\Class-CIT\BRB-ArrayTools Class\Pomeroy\ExpDescFile" and a "Browse" button.

At the bottom of the dialog are five buttons: "Back", "Next", "Cancel", "Reset", and "Help".

[Hands-on instructions]

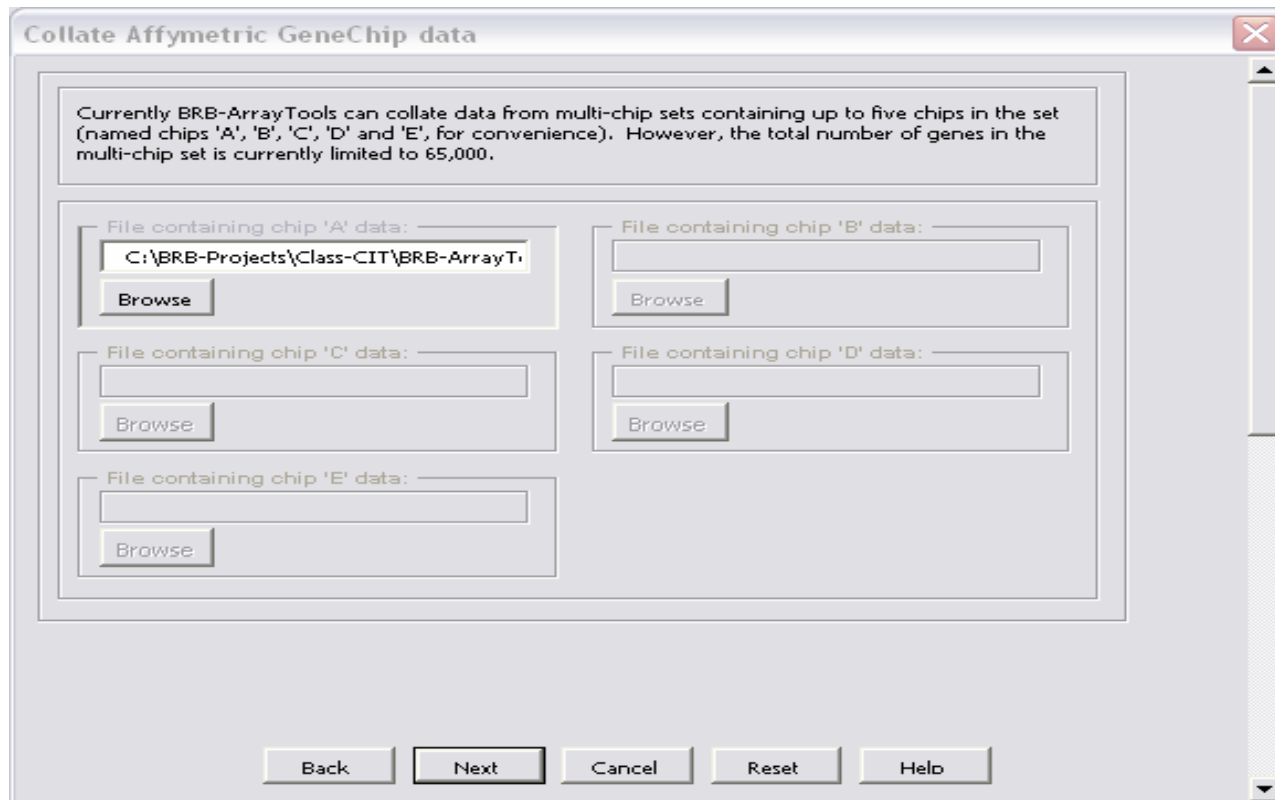
[Importing Pomeroy Data set]

- Select the option that the data is in a Horizontally aligned file.
- Choose **HuGeneFL** as the GeneChip set.
- Number of Chip types =1
- **Import annotations automatically**
- **Browse** for the following file which is also in the **Pomeroy** folder inside the **BRB-ArrayTools Class** folder which is on the **Desktop: ExpDescrMedulo.xls**

[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**: **Dataset_A2_multiple_tumor_samples.txt** and then click “Next”.



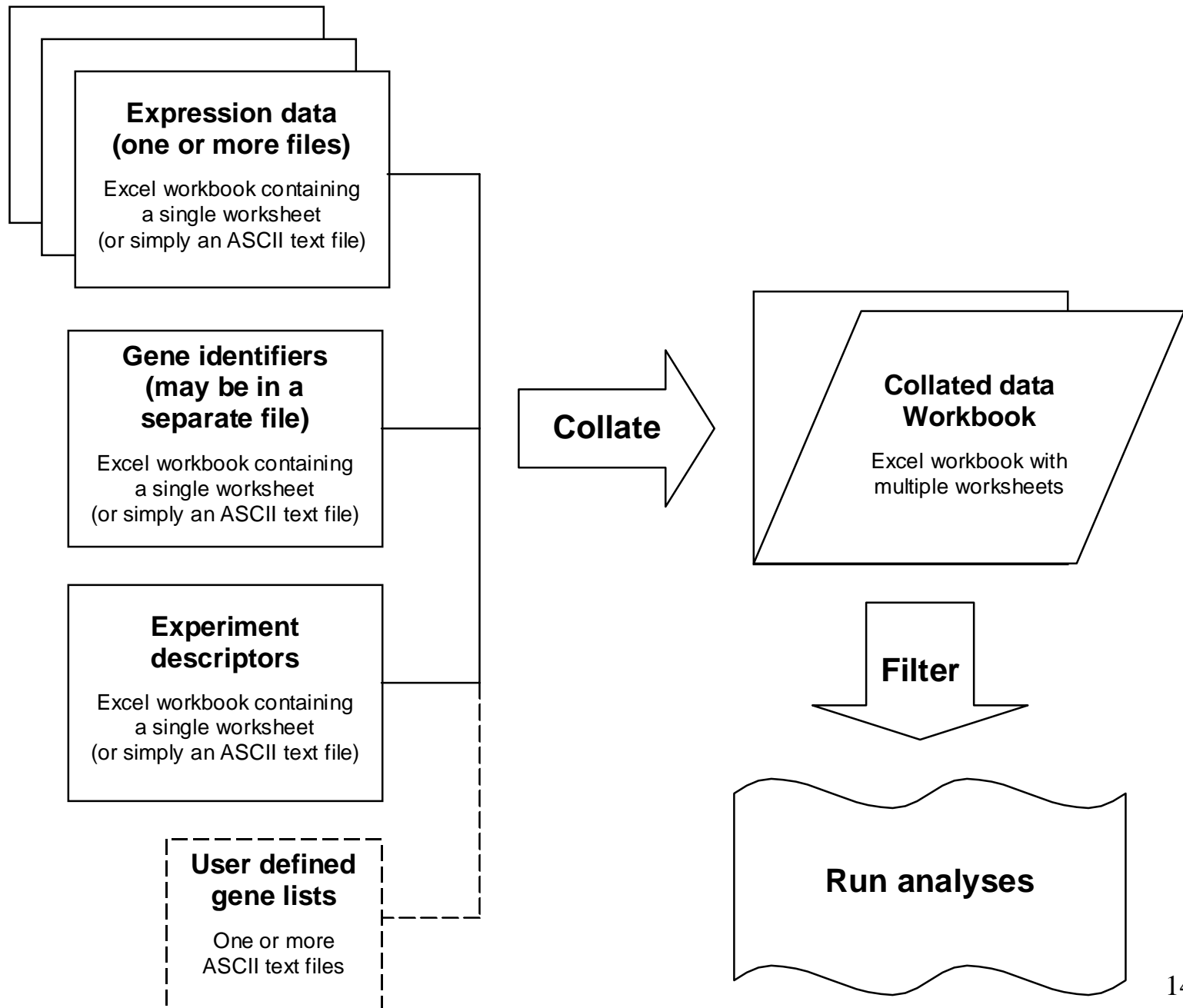
[Hands-on instructions]

[Importing Pomeroy Data set]

- Keep the defaults for Filtering.
- Save the Project in the folder “Pomeroy-Project”.
- The progress bar will indicate that the project is collating.
- Do you wish to annotate your data now? Please select “yes” and the “Utility” function to annotate the data will match the genes in the data set with pathways and defined genelists.

The collated project workbook

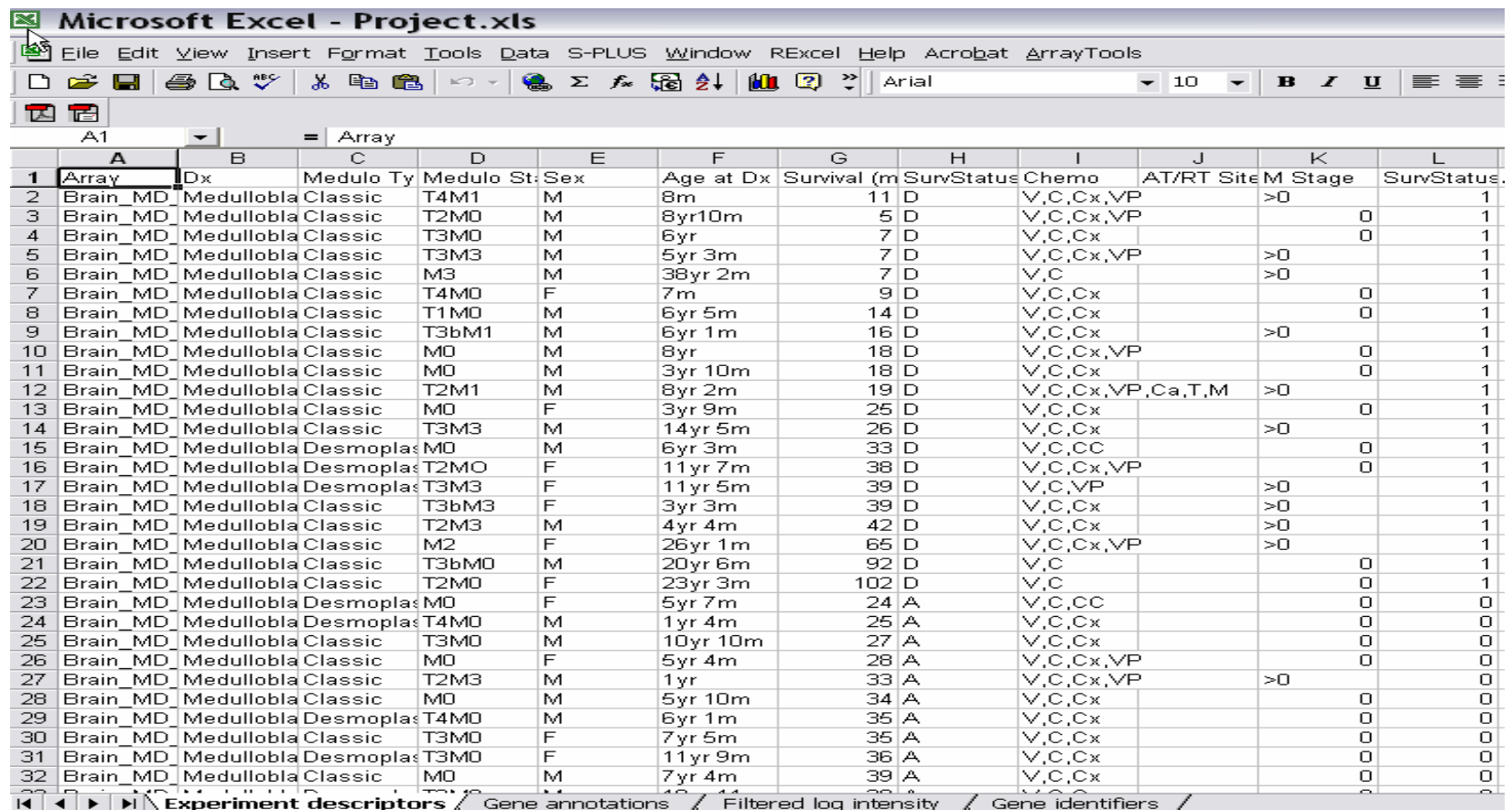
- This is the primary data object on which future analyses are run.
- Contains three primary worksheets:
 1. Experiment descriptors (may edit this to specify analyses)
 2. Gene identifiers
 3. Filtered log ratio (or Filtered log intensity)
- Additional results worksheets which may be automatically added:
 1. Gene annotations (obtained by running the menu item:
**Utilities → Annotate data →
Import Affymetrix or SOURCE annotations**)
 2. Scatterplot results
 3. Cluster analysis results



The collated project workbook

Experiment descriptor sheet

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



| 1 | Array | Dx | Medulo Ty | Medulo St | Sex | Age at Dx | Survival (m) | SurvStatus | Chemo | AT/RT Site | M Stage | SurvStatus |
|----|----------|------------|-----------|-----------|-----|-----------|--------------|------------|------------------|------------|---------|------------|
| 2 | Brain_MD | Medullobla | Classic | T4M1 | M | 8m | 11 D | | V,C,Cx,VP | | >0 | 1 |
| 3 | Brain_MD | Medullobla | Classic | T2M0 | M | 8yr10m | 5 D | | V,C,Cx,VP | | 0 | 1 |
| 4 | Brain_MD | Medullobla | Classic | T3M0 | M | 6yr | 7 D | | V,C,Cx | | 0 | 1 |
| 5 | Brain_MD | Medullobla | Classic | T3M3 | M | 5yr 3m | 7 D | | V,C,Cx,VP | | >0 | 1 |
| 6 | Brain_MD | Medullobla | Classic | M3 | M | 38yr 2m | 7 D | | V,C | | >0 | 1 |
| 7 | Brain_MD | Medullobla | Classic | T4M0 | F | 7m | 9 D | | V,C,Cx | | 0 | 1 |
| 8 | Brain_MD | Medullobla | Classic | T1M0 | M | 6yr 5m | 14 D | | V,C,Cx | | 0 | 1 |
| 9 | Brain_MD | Medullobla | Classic | T3bM1 | M | 6yr 1m | 16 D | | V,C,Cx | | >0 | 1 |
| 10 | Brain_MD | Medullobla | Classic | M0 | M | 8yr | 18 D | | V,C,Cx,VP | | 0 | 1 |
| 11 | Brain_MD | Medullobla | Classic | M0 | M | 3yr 10m | 18 D | | V,C,Cx | | 0 | 1 |
| 12 | Brain_MD | Medullobla | Classic | T2M1 | M | 8yr 2m | 19 D | | V,C,Cx,VP,Ca,T,M | | >0 | 1 |
| 13 | Brain_MD | Medullobla | Classic | M0 | F | 3yr 9m | 25 D | | V,C,Cx | | 0 | 1 |
| 14 | Brain_MD | Medullobla | Classic | T3M3 | M | 14yr 5m | 26 D | | V,C,Cx | | >0 | 1 |
| 15 | Brain_MD | Medullobla | Desmoplas | M0 | M | 6yr 3m | 33 D | | V,C,CC | | 0 | 1 |
| 16 | Brain_MD | Medullobla | Desmoplas | T2M0 | F | 11yr 7m | 38 D | | V,C,Cx,VP | | 0 | 1 |
| 17 | Brain_MD | Medullobla | Desmoplas | T3M3 | F | 11yr 5m | 39 D | | V,C,VP | | >0 | 1 |
| 18 | Brain_MD | Medullobla | Classic | T3bM3 | F | 3yr 3m | 39 D | | V,C,Cx | | >0 | 1 |
| 19 | Brain_MD | Medullobla | Classic | T2M3 | M | 4yr 4m | 42 D | | V,C,Cx | | >0 | 1 |
| 20 | Brain_MD | Medullobla | Classic | M2 | F | 26yr 1m | 65 D | | V,C,Cx,VP | | >0 | 1 |
| 21 | Brain_MD | Medullobla | Classic | T3bM0 | M | 20yr 6m | 92 D | | V,C | | 0 | 1 |
| 22 | Brain_MD | Medullobla | Classic | T2M0 | F | 23yr 3m | 102 D | | V,C | | 0 | 1 |
| 23 | Brain_MD | Medullobla | Desmoplas | M0 | F | 5yr 7m | 24 A | | V,C,CC | | 0 | 0 |
| 24 | Brain_MD | Medullobla | Desmoplas | T4M0 | M | 1yr 4m | 25 A | | V,C,Cx | | 0 | 0 |
| 25 | Brain_MD | Medullobla | Classic | T3M0 | M | 10yr 10m | 27 A | | V,C,Cx | | 0 | 0 |
| 26 | Brain_MD | Medullobla | Classic | M0 | F | 5yr 4m | 28 A | | V,C,Cx,VP | | 0 | 0 |
| 27 | Brain_MD | Medullobla | Classic | T2M3 | M | 1yr | 33 A | | V,C,Cx,VP | | >0 | 0 |
| 28 | Brain_MD | Medullobla | Classic | M0 | M | 5yr 10m | 34 A | | V,C,Cx | | 0 | 0 |
| 29 | Brain_MD | Medullobla | Desmoplas | T4M0 | M | 6yr 1m | 35 A | | V,C,Cx | | 0 | 0 |
| 30 | Brain_MD | Medullobla | Classic | T3M0 | F | 7yr 5m | 35 A | | V,C,Cx | | 0 | 0 |
| 31 | Brain_MD | Medullobla | Desmoplas | T3M0 | F | 11yr 9m | 36 A | | V,C,Cx | | 0 | 0 |
| 32 | Brain_MD | Medullobla | Classic | M0 | M | 7yr 4m | 39 A | | V,C,Cx | | 0 | 0 |

The collated project workbook

Gene identifier sheet

Contains gene identifiers provided by the user during collation.

| Microsoft Excel - Project.xls | | | | | |
|--------------------------------------------------------------------------------------|-----------|--------------------------------|--------------------------------------------|---------------------------------|--------|
| File Edit View Insert Format Tools Data S-PLUS Window RExcel Help Acrobat ArrayTools | | | | | |
| A1 = Probe set | | | | | |
| | A | B | C | D | E |
| | Probe set | Description | Gene symbol | Defined genelists | Filter |
| 5 | AB000220 | sema domain, immunoglobulin | SEMA3C | Perou's- Intrinsic- Breast-Canc | TRUE |
| 11 | AB000460 | chromosome 4 open reading fr | C4orf8 | | TRUE |
| 13 | AB000464 | chromosome 4 open reading fr | C4orf10 | | TRUE |
| 16 | AB000468 | ring finger protein 4 | RNF4 | | TRUE |
| 23 | AB001106 | glia maturation factor, beta | GMFB | | TRUE |
| 24 | AB001325 | aquaporin 3 | AQP3 | Perou's- Intrinsic- Breast-Canc | TRUE |
| 26 | AB002315 | KIAA0317 | KIAA0317 | | TRUE |
| 28 | AB002332 | clock homolog (mouse) | CLOCK | Circadian Rhythms | TRUE |
| 29 | AB002356 | MAP-kinase activating death d | MADD | TNFR1 Signaling Pathway | TRUE |
| 32 | AB002380 | Rho guanine nucleotide excha | ARHGEF12 | | TRUE |
| 33 | AB002382 | catenin (cadherin-associated p | CTNND1 | | TRUE |
| 37 | AB003102 | proteasome (prosome, macrop | PSMD11 | | TRUE |
| 38 | AB003103 | proteasome (prosome, macrop | PSMD12 | | TRUE |
| 39 | AB003177 | proteasome (prosome, macrop | PSMD9 | | TRUE |
| 40 | AB003698 | CDC7 cell division cycle 7 (S. | CDC7 | | TRUE |
| 41 | AB004884 | tousled-like kinase 2 | TLK2 | Phosphatidylinositol signaling | TRUE |
| 50 | AC000064 | GATA zinc finger domain conta | GATAD1 | | TRUE |
| 55 | AC002045 | nuclear pore complex interacti | NP1P /// LOC283970 /// LOC339047 /// LOC39 | | TRUE |
| 56 | AC002045 | | Multiple Gene Symbols | | TRUE |
| 61 | AC002115 | cytochrome c oxidase subunit | COX6B1 | Oxidative phosphorylation | TRUE |
| 96 | AF005775 | CASP8 and FADD-like apoptos | CFLAR | FAS signaling pathway (CD95 | TRUE |
| 104 | AF007875 | dolichyl-phosphate mannosyltr | DPM1 | N-Glycan biosynthesis | TRUE |
| 106 | AF008937 | syntaxin 16 | STX16 | | TRUE |
| 107 | AF009301 | membrane-associated ring fing | MARCH6 | | TRUE |
| 108 | AF009368 | cAMP responsive element bind | CREB3 | | TRUE |
| 111 | AF010193 | SMAD, mothers against DPP | SMAD7 | | TRUE |
| 116 | AF015913 | SKB1 homolog (S. pombe) | SKB1 | | TRUE |
| 139 | AFFX-HSA | actin, beta | ACTB | Chromatin Remodeling by hSV | TRUE |
| 140 | AFFX-HSA | actin, beta | ACTB | Chromatin Remodeling by hSV | TRUE |
| 141 | AFFX-HSA | actin, beta | ACTB | Chromatin Remodeling by hSV | TRUE |
| 142 | AFFX-HSA | actin, beta | ACTB | Chromatin Remodeling by hSV | TRUE |

The collated project workbook

Filtered log ratio or log intensity sheet

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

Microsoft Excel - Project.xls

File Edit View Insert Format Tools Data S-PLUS Window RExcel Help Acrobat ArrayTools

B2 = 0

| | A | B | C | D | E | F | G | H |
|-----|-------------------------------------------|---------|---------|------|----------|--------------|---------------|--------|
| | Click to display the data | Missing | P-Value | Rank | Variance | Num 1.5-Fold | Absent (Affy) | Filter |
| 1 | Probe set | | | | | | | |
| 5 | AB000220_at | 0 | | | | 64 | 34 | TRUE |
| 11 | AB000460_at | 0 | | | | 37 | 1 | TRUE |
| 13 | AB000464_at | 0 | | | | 55 | 27 | TRUE |
| 16 | AB000468_at | 0 | | | | 55 | 15 | TRUE |
| 23 | AB001106_at | 0 | | | | 56 | 14 | TRUE |
| 24 | AB001325_at | 0 | | | | 41 | 12 | TRUE |
| 26 | AB002315_at | 0 | | | | 56 | 36 | TRUE |
| 28 | AB002332_at | 0 | | | | 56 | 22 | TRUE |
| 29 | AB002356_s_at | 0 | | | | 45 | 12 | TRUE |
| 32 | AB002380_at | 0 | | | | 73 | 23 | TRUE |
| 33 | AB002382_at | 0 | | | | 49 | 32 | TRUE |
| 37 | AB003102_at | 0 | | | | 53 | 7 | TRUE |
| 38 | AB003103_at | 0 | | | | 57 | 18 | TRUE |
| 39 | AB003177_at | 0 | | | | 49 | 26 | TRUE |
| 40 | AB003698_at | 0 | | | | 55 | 39 | TRUE |
| 41 | AB004884_at | 0 | | | | 49 | 26 | TRUE |
| 50 | AC000064_cds1_at | 0 | | | | 47 | 28 | TRUE |
| 55 | AC002045_xpt1_at | 0 | | | | 56 | 27 | TRUE |
| 56 | AC002045_xpt2_s_at | 0 | | | | 47 | 0 | TRUE |
| 61 | AC002115_cds1_at | 0 | | | | 36 | 4 | TRUE |
| 96 | AF005775_at | 0 | | | | 33 | 31 | TRUE |
| 104 | AF007875_at | 0 | | | | 55 | 35 | TRUE |
| 106 | AF008937_at | 0 | | | | 49 | 33 | TRUE |
| 107 | AF009301_at | 0 | | | | 55 | 20 | TRUE |
| 108 | AF009368_at | 0 | | | | 39 | 16 | TRUE |
| 111 | AF010193_at | 0 | | | | 56 | 31 | TRUE |
| 116 | AF015913_at | 0 | | | | 49 | 21 | TRUE |
| 139 | AFFX-HSAC07/X00351_3_at | 0 | | | | 26 | 0 | TRUE |
| 140 | AFFX-HSAC07/X00351_3_st | 0 | | | | 55 | 5 | TRUE |

Experiment descriptors / Gene annotations / **Filtered log intensity** / Gene identifiers /

Filter Mode

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.

| Microsoft Excel - Project.xls | | | | | | | | | | |
|--------------------------------------------------------------------------------------|--------------------------|-------------|----------|-----------|------------|-------------|------------|-------------|--------------|--------|
| File Edit View Insert Format Tools Data S-PLUS Window RExcel Help Acrobat ArrayTools | | | | | | | | | | |
| A1 = Probe set (Double-click) | | | | | | | | | | |
| | A | B | C | D | E | F | G | H | I | J |
| 1 | Probe set (Double-click) | Name | Gb acc | UGCluster | Symbol | LLID | Chromosome | Cytoband | GO | Filter |
| 5 | AB000220 | sema dom | AB000220 | Hs.269109 | SEMA3C | 10512 | 7 | Chr:7q21-c | | TRUE |
| 11 | AB000460 | chromosom | AB000460 | Hs.125348 | C4orf8 | 8603 | 4 | Chr:4p16.3 | | TRUE |
| 13 | AB000464 | chromosom | AB000464 | Hs.125348 | C4orf10 | 317648 | 4 | Chr:4p16.3 | | TRUE |
| 16 | AB000468 | ring finger | AB000468 | Hs.66394 | RNF4 | 6047 | 4 | Chr:4p16.3 | cellular co | TRUE |
| 23 | AB001106 | glia matur | AB001106 | Hs.151413 | GMFB | 2764 | 14 | Chr:14q22 | cellular co | TRUE |
| 24 | AB001325 | aquaporin | AB001325 | Hs.234642 | AQP3 | 360 | 9 | Chr:9p13 | ##### | TRUE |
| 26 | AB002315 | KIAA0317 | AB002315 | Hs.497417 | KIAA0317 | 9870 | 14 | Chr:14q24 | ##### | TRUE |
| 28 | AB002332 | clock hom | AB002332 | Hs.436975 | CLOCK | 9575 | 4 | Chr:4q12 | cellular co | TRUE |
| 29 | AB002356 | MAP-kinas | AB002356 | Hs.82548 | MADD | 8567 | 11 | Chr:11p11 | ##### | TRUE |
| 32 | AB002380 | Rho guanir | AB002380 | Hs.24598 | ARHGEF1 | 23365 | 11 | Chr:11q23 | molecular | TRUE |
| 33 | AB002382 | catenin (ca | AB002382 | Hs.166011 | CTNND1 | 1500 | 11 | Chr:11q11 | ##### | TRUE |
| 37 | AB003102 | proteasom | AB003102 | Hs.443379 | PSMD11 | 5717 | 17 | Chr:17q11 | cellular co | TRUE |
| 38 | AB003103 | proteasom | AB003103 | Hs.4295 | PSMD12 | 5718 | 17 | Chr:17q24 | cellular co | TRUE |
| 39 | AB003177 | proteasom | AB003177 | Hs.131151 | PSMD9 | 5715 | 12 | Chr:12q24 | ##### | TRUE |
| 40 | AB003698 | CDC7 cell | AB003698 | Hs.533573 | CDC7 | 8317 | 1 | Chr:1p22 | cellular co | TRUE |
| 41 | AB004884 | tousled-lik | AB004884 | Hs.445078 | TLK2 | 11011 | 17 | Chr:17q23 | cellular co | TRUE |
| 50 | AC000064 | GATA zinc | AC000064 | Hs.21145 | GATAD1 | 57798 | 7 | Chr:7q21-c | biological p | TRUE |
| 55 | AC002045 | nuclear po | AC002045 | Hs.446747 | NP1P /// L | 283970 /// | 16 | Chr:16p13 | biological p | TRUE |
| 56 | AC002045 | | AC002045 | Hs.558978 | Multiple G | 23117 /// 2 | 16 | Chr:16p13 | | TRUE |
| 61 | AC002115 | cytochrom | AC002115 | Hs.431668 | COX6B1 | 1340 | 19 | Chr:19q13 | ##### | TRUE |
| 96 | AF005775 | CASP8 an | AF005775 | Hs.390736 | CFLAR | 8837 | 2 | Chr:2q33-c | | TRUE |
| 104 | AF007875 | dolichyl-ph | AF007875 | Hs.301898 | DPM1 | 8813 | 20 | Chr:20q13 | biological p | TRUE |
| 106 | AF008937 | syntaxin 11 | AF008937 | Hs.307913 | STX16 | 8675 | 20 | Chr:20q13 | molecular | TRUE |
| 107 | AF009301 | membrane | AF009301 | Hs.432862 | MARCH6 | 10299 | 5 | Chr:5p15.2 | ##### | TRUE |
| 108 | AF009368 | cAMP resp | AF009368 | Hs.522110 | CREB3 | 10488 | 9 | Chr:9pter-p | cellular co | TRUE |
| 111 | AF010193 | SMAD, mo | AF010193 | Hs.465087 | SMAD7 | 4092 | 18 | Chr:18q21 | cellular co | TRUE |
| 116 | AF015913 | SKB1 hom | AF015913 | Hs.367854 | SKB1 | 10419 | 14 | Chr:14q11 | ##### | TRUE |
| 139 | AFX-HSA | actin, beta | X00351 | Hs.520640 | ACTB | 60 | 7 | Chr:7p15-p | cellular co | TRUE |
| 140 | AFX-HSA | actin, beta | X00351 | Hs.520640 | ACTB | 60 | 7 | Chr:7p15-p | cellular co | TRUE |
| 141 | AFX-HSA | actin, beta | X00351 | Hs.520640 | ACTB | 60 | 7 | Chr:7p15-p | cellular co | TRUE |
| 142 | AFX-HSA | actin, beta | X00351 | Hs.520640 | ACTB | 60 | 7 | Chr:7p15-p | cellular co | TRUE |

Part III:

Data filtering and normalization options

[Hands-on instructions]

[Data filtering-Pomeroy]

1. Click on **ArrayTools** → **Filter** and **subset the data**.
2. Click on the three buttons **Spot filters**, **Normalization**, and **Gene filters** at the TOP of the form, to see the available options and view the current settings applied on the dataset.
3. Under Normalization tab select **Normalize (center) each array** using **median array** as reference.
4. Now click **OK** to apply the default filtering settings onto the data.

Data filtering options

Dual-Channel: Spot filter

- Intensity filter: May filter out spots with low intensity in dual channels, or threshold the low-intensity channel in forming the log-ratio.
- Spot flag filter: May filter out spots with flag values outside a specified range, or spots with flags containing specified values
- Spot size filter: May exclude a spot which is computed from too few pixels

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”.

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: Option 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)

Normalization-Dual Channel

For dual-channel data:

- Options to median-center all arrays to 0, center arrays so that median over housekeeping genes is 0, or adjust log-ratios based on intensity levels (using lowess smoother).
- Print-tip Group(Block): The option to center using the median for each print-tip group or using the lowess smoother within each print-tip group.

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

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 genelists 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.

Part IV:

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
- Click on plots to hyperlink to clone reports
- Double-click to view gene annotations (if available)

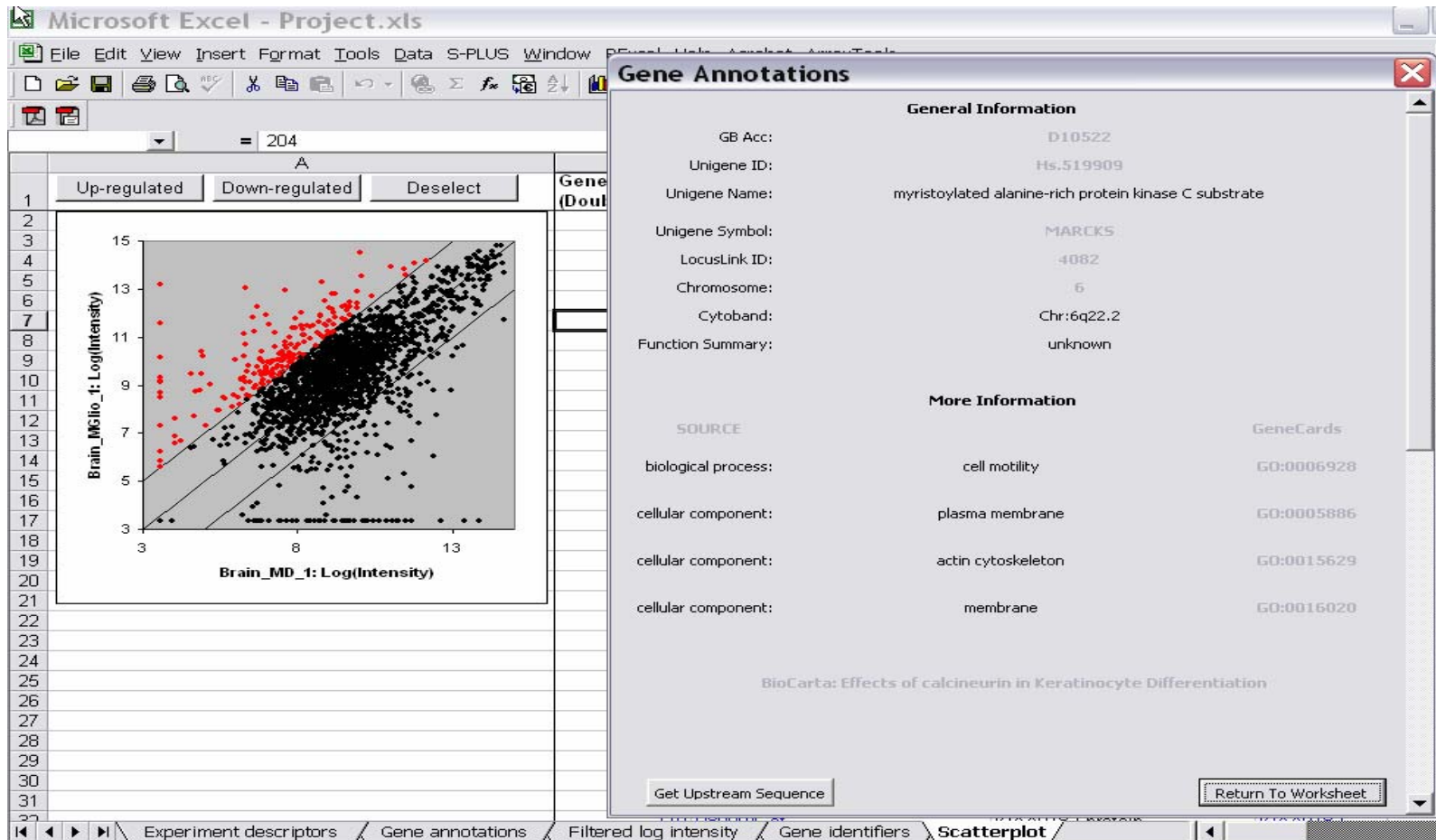
[Hands-on instructions]

[Scatterplot of experiment v. experiment-Pomeroy Data]

1. Click on **ArrayTools** → **Scatterplot** → **Experiment v. Experiment**.
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. For the **Outlier lines**, set fold-difference parameter as 4. Then click **OK**.
3. When the plot appears, click on any point in the scatterplot to identify the selected point in the workspace to the right of the plot.
4. Try clicking the **Up-regulated** and **Down-regulated** buttons to select the points which fall outside the outlier lines (i.e., points for which the x- and y-values in the unlogged scale have a fold-difference greater than 4).

Scatterplot of experiment v. experiment

Pomeroy data – plot of log-intensity vs log-intensity showing genes differentially expressed between two experiments.



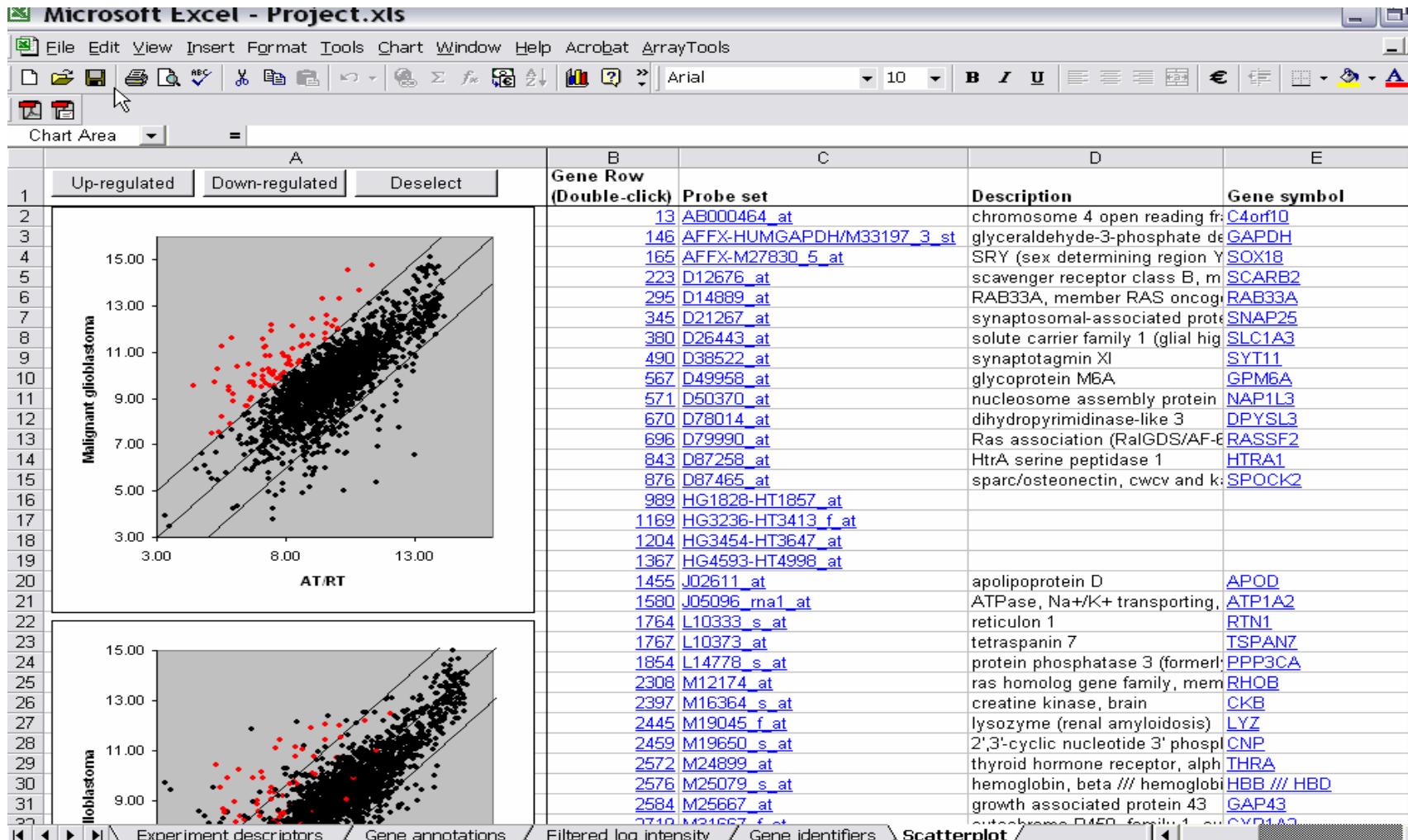
[Hands-on instructions]

[Scatterplot of phenotype averages]

1. Now click on **ArrayTools** → **Scatterplot** → **Phenotype averages**.
2. Select the **Dx** variable as the phenotype class to average over, and Fold-difference to 4, then click **OK**.
3. When the scatterplots appear, if you scroll down, you will see that there are ten plots, comparing the five different categories.
4. If you select the AT/RT versus Malignant glioblastoma plot (the first plot) and click the **Up-regulated** button, you will see all the genes for which the geometric mean of the AT/RT arrays is at least 4-times the geometric mean of the Malignant glioblastoma arrays. These genes are also highlighted in the other plots as well.

Scatterplot of phenotype averages

Pomeroy dataset



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
- Class Comparison Between Red and Green Channel.
- Gene Set Expression Comparison.
- ANOVA models

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.

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 tool

Between red and green channels

- Similar to the other class comparison tool, but compares between red and green samples within arrays, rather than between the red samples on groups of arrays.
- Used for Block Design, or for a Reference Design where the comparison of interest is between the test samples and reference sample rather than between test samples.
- Example of situation where this tool might be used:
 - Array1 contains (red=test1, green=control1)
 - Array2 contains (red=test2, green=control2)
 - Array3 contains (red=control1, green=test1), etc.

- Experiment descriptors:

| RedID | GreenID | RedClass | GreenClass |
|-------|---------|----------|------------|
| T1 | C1 | Test | Control |
| T2 | C2 | Test | Control |
| C1 | T1 | Control | Test |

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

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.

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.

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 **Dx** 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. 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

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 **Dx** variable as the column defining the classes, select the **90th percentile** option, and leave all other parameters at default levels. Now click **OK** to exit the options dialog, and click **OK** on the main dialog to launch the analysis.

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.

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.
- For each group of gene that is tested for significance, a summary measure of the univariate p-values of the genes within that group is compared against summary measures obtained from a random sampling of groups of genes of the same size. A permutation p-value for the group of genes is obtained by ranking its actual p-value compared to the random sampling.

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
 - Kolmogorov-Smirnov statistic.

Hotelling's T-square test

- Tests the hypothesis that no genes in the gene set are significant.
- Principal components are computed for all the genes in the gene set using all the arrays.
- Hotelling's T-sq is a multivariate test of the univariate t- and f- tests for testing the class means.
- Uses a large sample approximation (min # of arrays =20)

[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 **Dx** 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.

Exercise-I

Using the Perou 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 control for false discoveries obtain another genelist with a 95% confidence level and controlling for 40% False discoveries.**
- ? Explain the differences between the two outputs.**
- ? Using the common set of significant genes obtained from the above two analyses, run a scatterplot of phenotype averages with 2 fold difference and comment on the downward regulated genes.**

Components of Class Prediction

- Feature(gene) selection
 - which genes will be included in the model.
- Select model type.
 - choose prediction method (DLDA,CCP etc)
- Fit the parameters for the model.
- Evaluating the Classifier
 - Cross-validation

Class prediction tool

- 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)
- 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.
- In addition, we have added the option to select genes using “gene pairs” by the “greedy pair” method –Bo & Jonassen

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

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 $\frac{1}{2}$ - 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 Shrunk 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 δ and the number of genes included in the classifier is determined by the value of δ .

[Hands-on instructions]

[Class prediction –Pomeroy data]

1. Run the Class Prediction tool by clicking on **ArrayTools** → **Class prediction** → **Class prediction**.
2. Select the **Dx** 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. De select the Compound Covariate Predictor,BCCP and SVM.
5. Select **Options**, check the box for **Use separate test set**, and select the column “TrainingSet”.
6. Leave all other options at default levels, and click **OK**.
7. Note the Array Ids which have been misclassified by all methods.

Survival Risk Prediction

- Develops a gene expression based predictor of survival risk groups allowing for up to 3 covariates in the model.
- Selects genes that correlate with survival based on a p-value threshold.
- First k-principal components are computed.
- A k-variable Cox proportional hazards regression is performed with k pc's as predictors.

Survival Risk Group Prediction

- Cross-validate using LOOCV or K-fold
- Develop supervised pc PH model for training set.
- Compute cross-validated predictive index for i^{th} case using the PH model for training set.
- Compute the predictive risk percentile index for i^{th} case among the predictive indices for the cases in the training set.
- Plot K-M survival curves.
- Compute the log-rank statistic comparing the cross-validated K-M curves.

[Hands-on instructions]

Survival Risk Prediction

- Using the Pomeroy data set, clicking on **ArrayTools** → **Survival Analysis** → **Survival Risk prediction**.
- Select “**SurvStatusCode**” for the survival status column and “**Survival(months)**” as the survival time column.
- Keeping all other options as defaults. Select the “**Age**” as a clinical covariate.
- Now click **OK** on the main dialog to launch the analysis.

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 **Dx** 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 100 to 200 and arrays 30 to 60 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** → **Multidimensional scaling** → **of samples**.
2. Now click on the **Options** button, and choose **Dx** as the variable to **Color the rotating scatterplot**. Click **OK** to exit the options dialog, and 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.

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

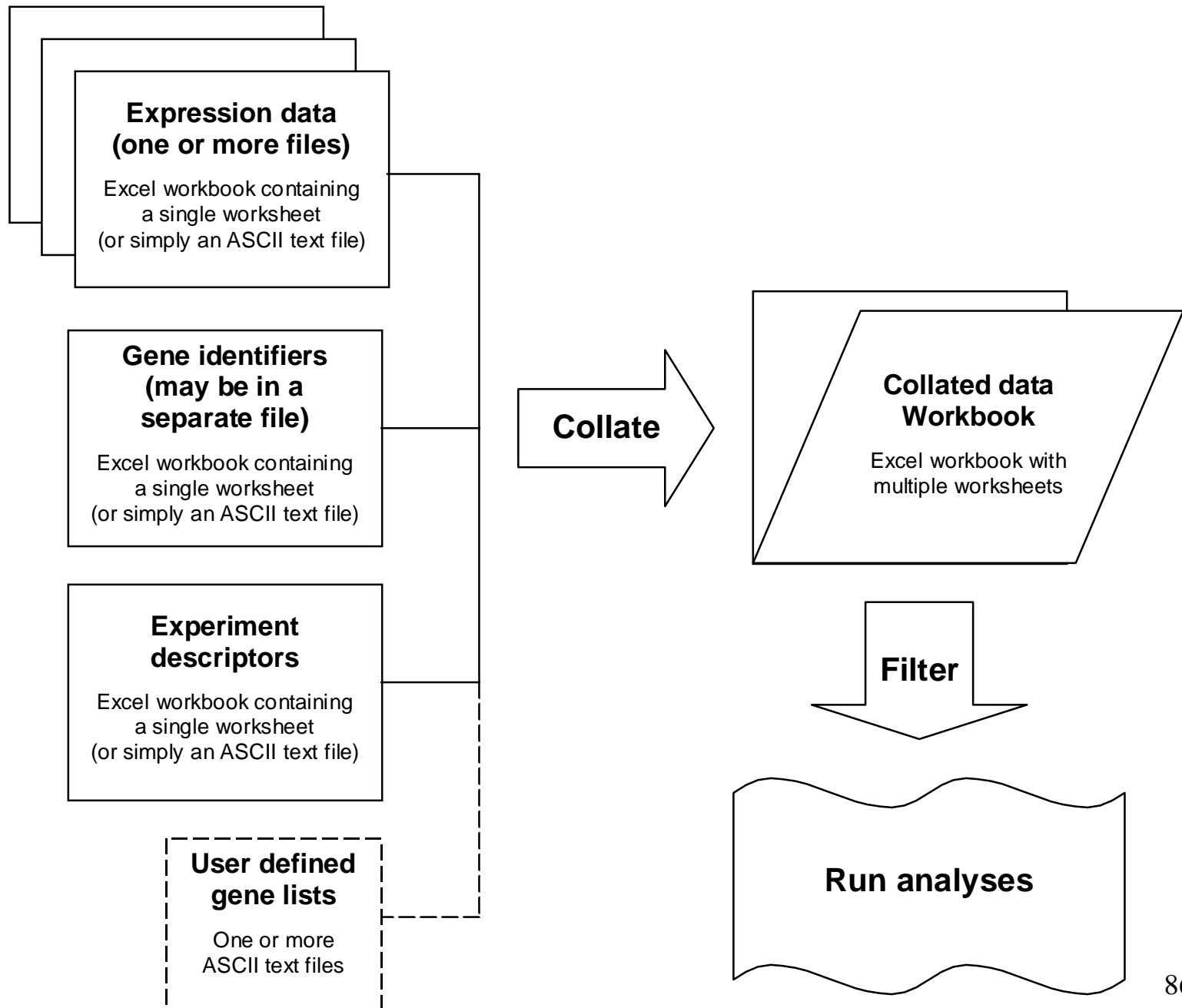
- Analysis of Variance – Up to four-way ANOVA. Options to include blocking factors or use random variance model.
- ANOVA of log intensities – For dual-channel non-reference designs, model includes gene-specific array effect, dye effect, and class effect. Option to use random variance model.
- ANOVA for 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.

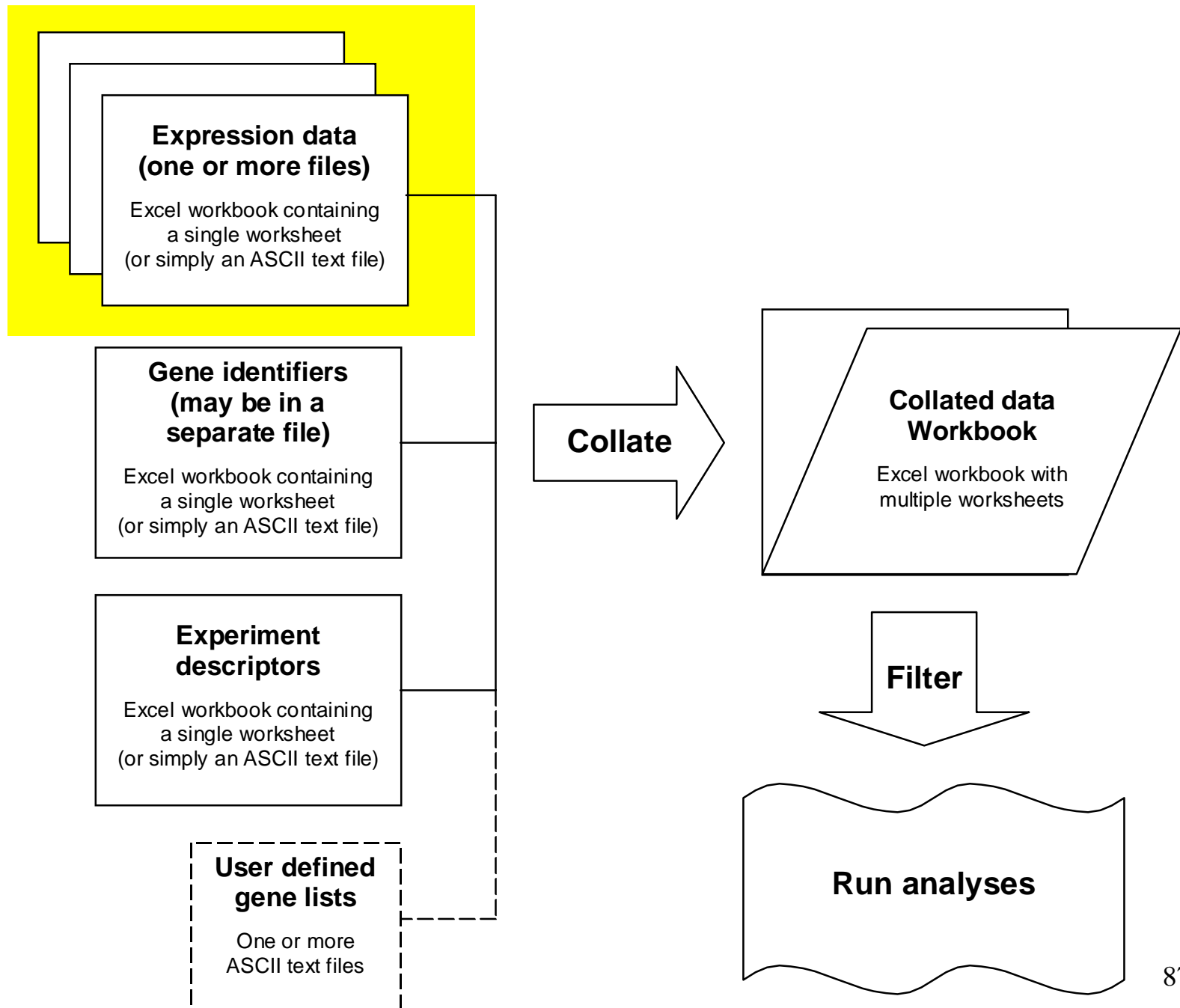
[Additional Plugins]

- 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.

Part V:

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





Expression data

- Input data as tab-delimited ASCII files (or Excel spreadsheets) in one of the following three formats:
 1. Horizontally aligned
 2. Separate files
 3. Multi-chip sets
- 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.
- For Affymetrix data, expression data files should be PROBESET-level data if using the Data Import Wizard. Affymetrix CEL files should be imported using a specialized utility included with BRB-ArrayTools.

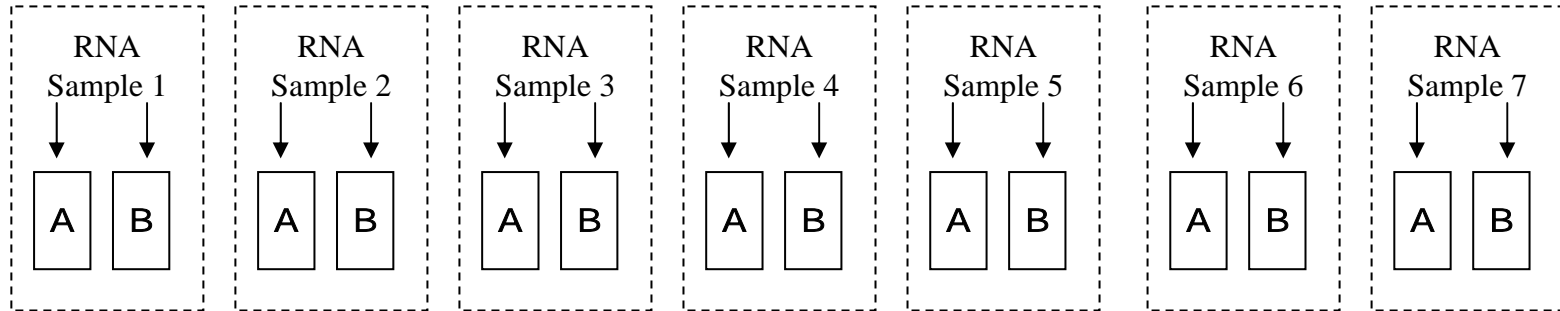
Expression data

Horizontally aligned data example

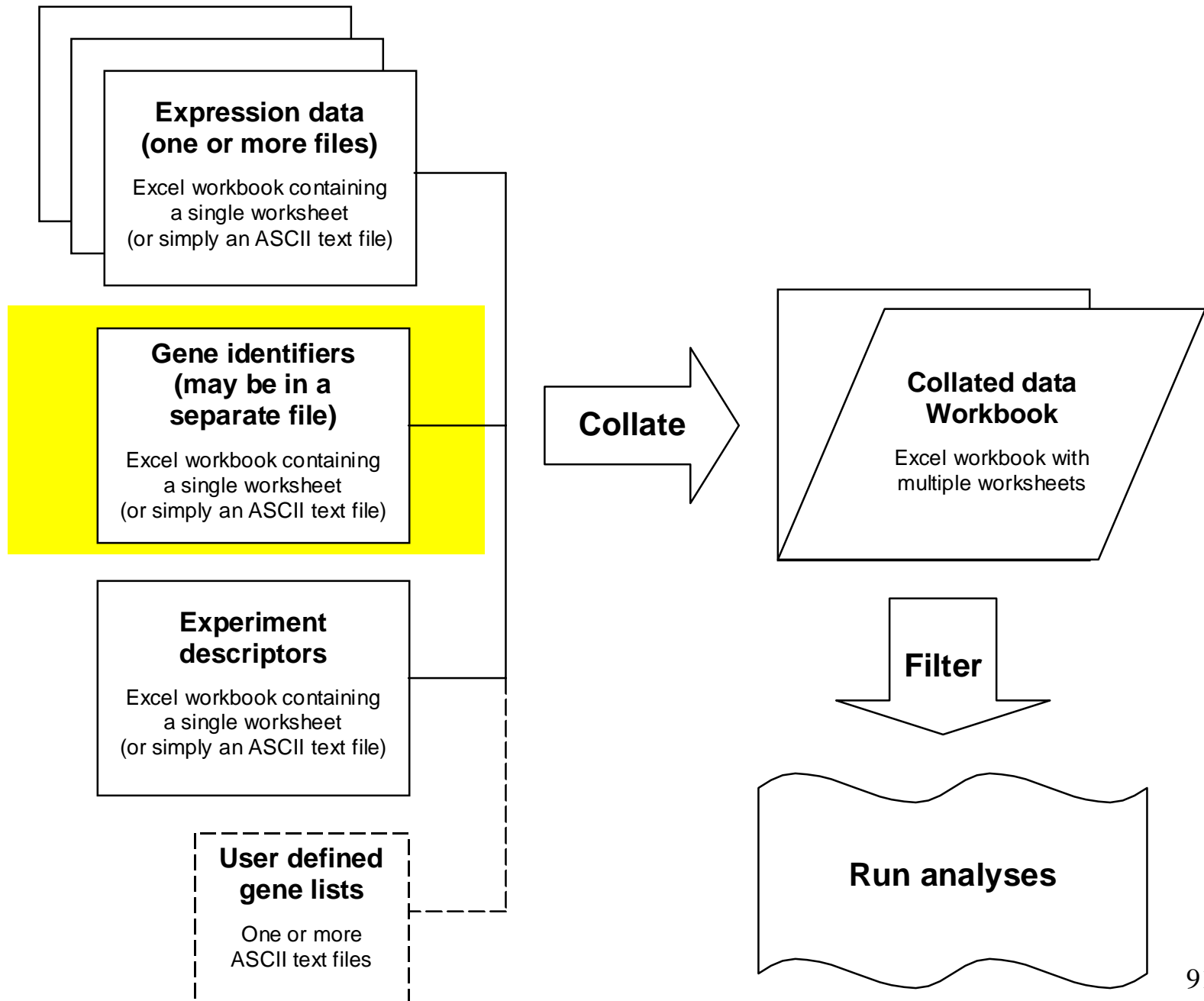
| | | | | Array data block #1 | | | Array data block #2 | | | Array data block #3 | | |
|----|--------|--------------|-------------------------|---------------------|---------|--------|---------------------|---------|--------|---------------------|---------|--------|
| | | | | | | | | | | | | |
| | A | B | C | D | E | F | G | H | I | J | K | L |
| 1 | Wellid | Clone | Description | Red_1 | Green_1 | Flag_1 | Red_2 | Green_2 | Flag_2 | Red_3 | Green_3 | Flag_3 |
| 2 | 600001 | IMAGE:604856 | adhesion selectin B Mm | 21363 | 13268 | 0 | 19674 | 11840 | 0 | 11938 | 4870 | |
| 3 | 600002 | IMAGE:619876 | adhesion VCAM-1 Mm | 16895 | 11908 | 0 | 45073 | 30279 | 0 | 16194 | 7591 | |
| 4 | 600003 | IMAGE:442991 | adhesion ELAM Mm.21 | 3823 | 2511 | 0 | 8238 | 3657 | 0 | 6574 | 1962 | |
| 5 | 600004 | IMAGE:615729 | adhesion integrinB-6 | 11277 | 5950 | 0 | 11045 | 6706 | 0 | 7020 | 3879 | |
| 6 | 600005 | IMAGE:522319 | adhesion integrin a5 Mm | 8979 | 3402 | 0 | 12431 | 3497 | 0 | 7650 | 1871 | |
| 7 | 600006 | IMAGE:576194 | adhesion integrin B1 | 17472 | 12238 | 0 | 14281 | 10961 | 0 | 14337 | 6918 | |
| 8 | 600007 | IMAGE:533853 | adhesion thrombosporin | 14204 | 8937 | 0 | 14476 | 4305 | 0 | 9043 | 2321 | |
| 9 | 600008 | IMAGE:476523 | adhesion ICAM Mm.394 | 17872 | 9822 | 0 | 22568 | 12239 | 0 | 11049 | 5572 | |
| 10 | 600009 | IMAGE:538626 | adhesion integrin a4 Mm | 35025 | 15216 | 0 | 43500 | 14654 | 0 | 19379 | 5698 | |
| 11 | 600010 | IMAGE:478744 | adhesion integrin a2 | 18122 | 9274 | 0 | 21378 | 10640 | 0 | 12177 | 4697 | |
| 12 | 600011 | IMAGE:679592 | adhesion integrin B8 Mr | 49522 | 25469 | 0 | 53653 | 21495 | 0 | 30237 | 8461 | |
| 13 | 600012 | IMAGE:426454 | adhesion integrin B7 Mr | 38276 | 17583 | 0 | 40191 | 15761 | 0 | 21316 | 6757 | |
| 14 | 600013 | IMAGE:573223 | adhesion integrin a6 | 2697 | 1604 | 0 | 2400 | 984 | 0 | 1473 | 579 | |
| 15 | 600014 | IMAGE:537501 | adhesion desmoplakin I | 8862 | 5660 | 0 | 11860 | 7598 | 0 | 7032 | 2228 | |
| 16 | 600015 | IMAGE:443962 | adhesion junction plak. | 5272 | 5945 | 0 | 5140 | 3944 | 0 | 2023 | 1335 | |
| 17 | 600016 | IMAGE:639320 | adhesion selectin P | 3813 | 3368 | 0 | 4176 | 3991 | 0 | 3841 | 2332 | |
| 18 | 600017 | IMAGE:677203 | adhesion selectin E Mm | 5201 | 3209 | 0 | 5314 | 2058 | 0 | 2305 | 709 | |
| 19 | 600018 | IMAGE:672927 | adhesion SQM1 | 8793 | 4038 | 0 | 13467 | 4856 | 0 | 7651 | 1788 | |
| 20 | 600019 | IMAGE:535792 | adhesion cadherin 5 Mm | 9162 | 15130 | 0 | 7701 | 12335 | 0 | 3214 | 5331 | |
| 21 | 600020 | IMAGE:473150 | adhesion thrombospond | 16010 | 5794 | 0 | 20450 | 7963 | 0 | 10764 | 3165 | |
| 22 | 600021 | IMAGE:639878 | adhesion integrin a9 | 3649 | 3065 | 0 | 4291 | 3198 | 0 | 1911 | 1383 | |
| 23 | 600022 | IMAGE:521884 | adhesion fibronectin | 3115 | 2737 | 0 | 7156 | 7223 | 0 | 6637 | 1858 | |
| 24 | 600023 | MP:1B11 | adhesion integrin B1 | 3139 | 1770 | 0 | 2900 | 822 | 0 | 1417 | 505 | |

Expression data

Multi-chip sets



- Total gene set is divided among a set of up to 5 arrays, labeled chip types 'A', 'B', 'C', 'D', and 'E' for convenience
- The same RNA sample is hybridized to all chips in the set
- Multi-chip sets may be in horizontally aligned files (one file per chip type), or separate files (one expression data folder per chip type)



Gene identifiers

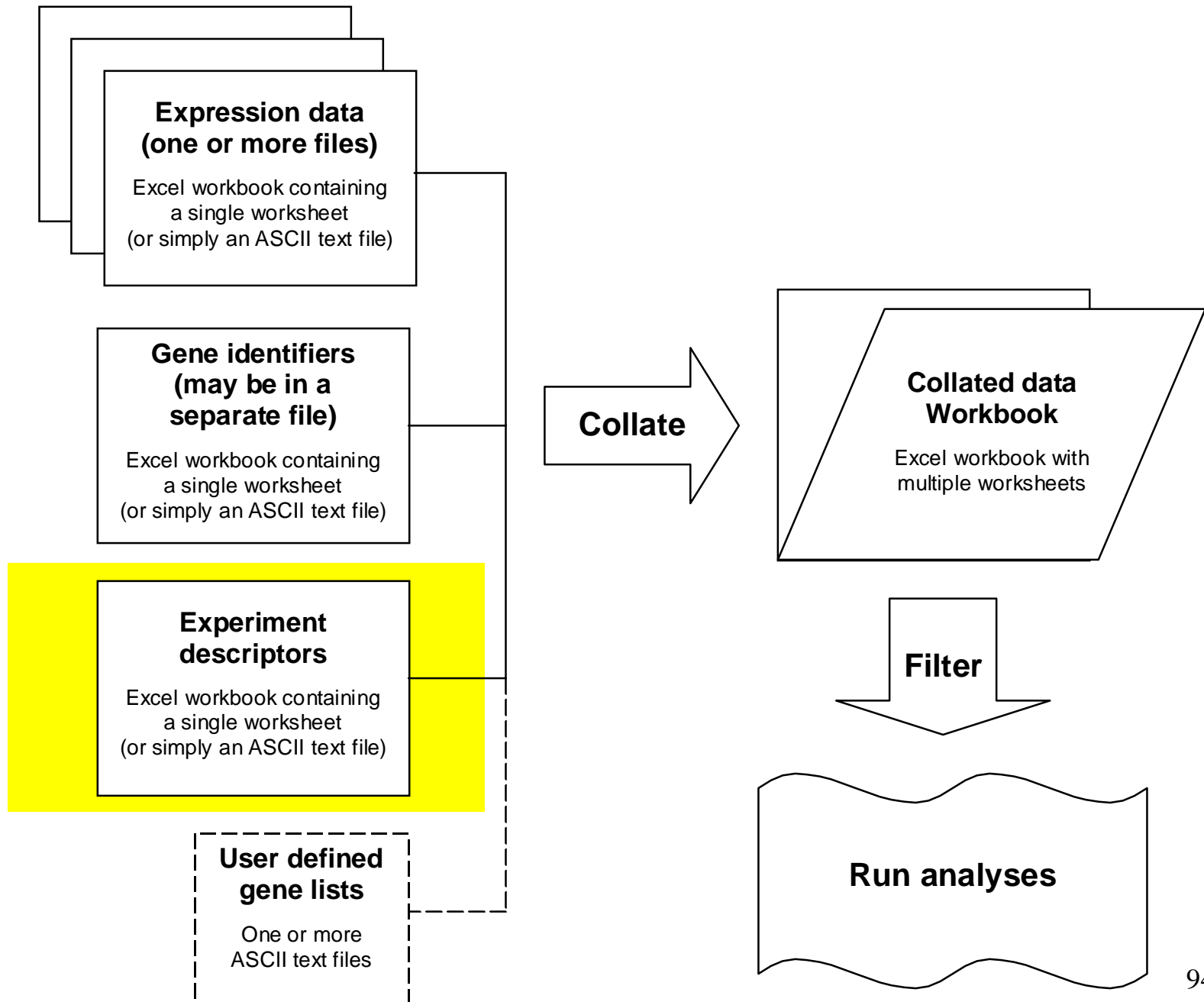
- A gene identifiers file is optional, but highly recommended for annotation purposes.
- Gene identifiers which may be used for hyperlinking are: clone ids, UniGene cluster id or gene symbol, GenBank accessions, and probe set ids.

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 | | | | | | | |

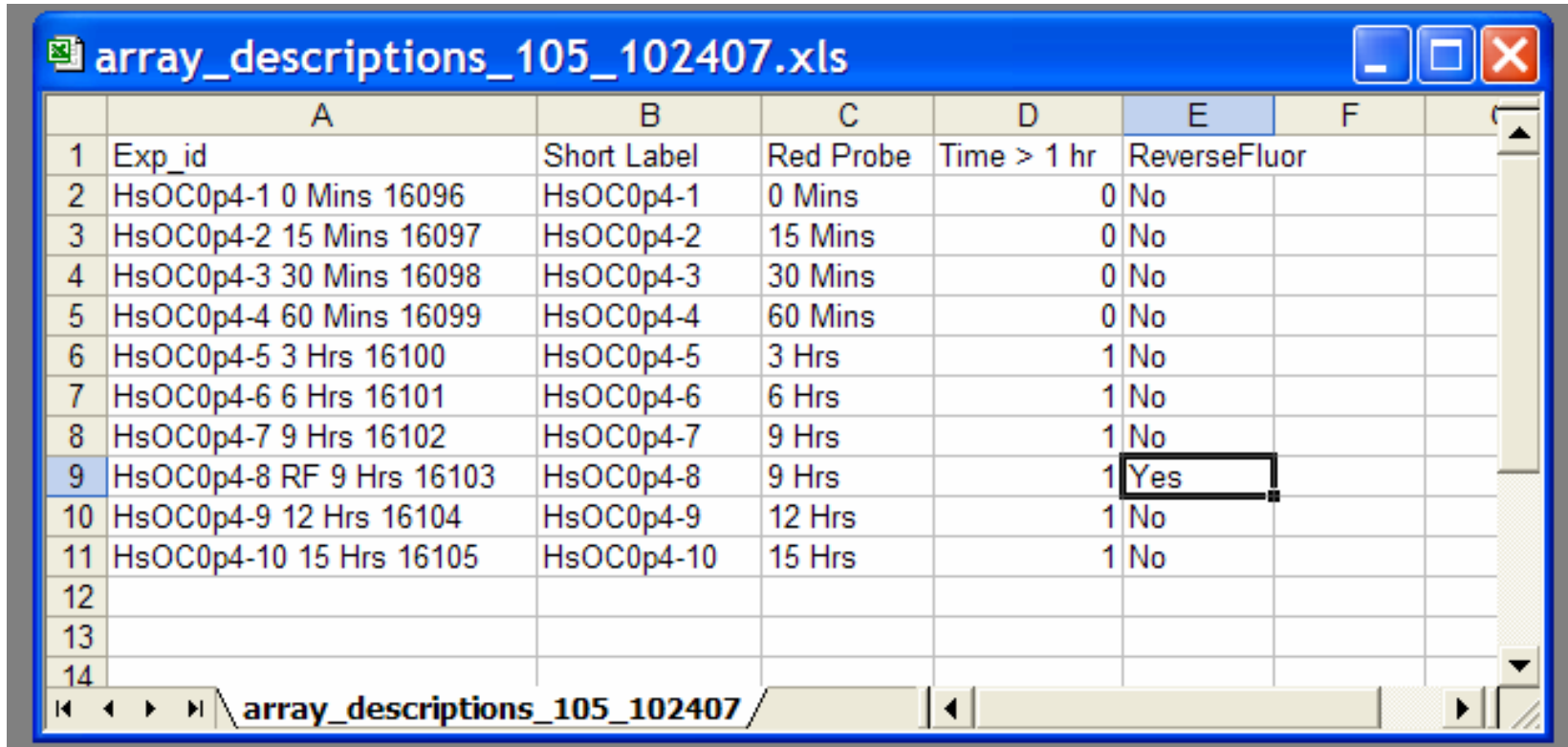


Experiment descriptors

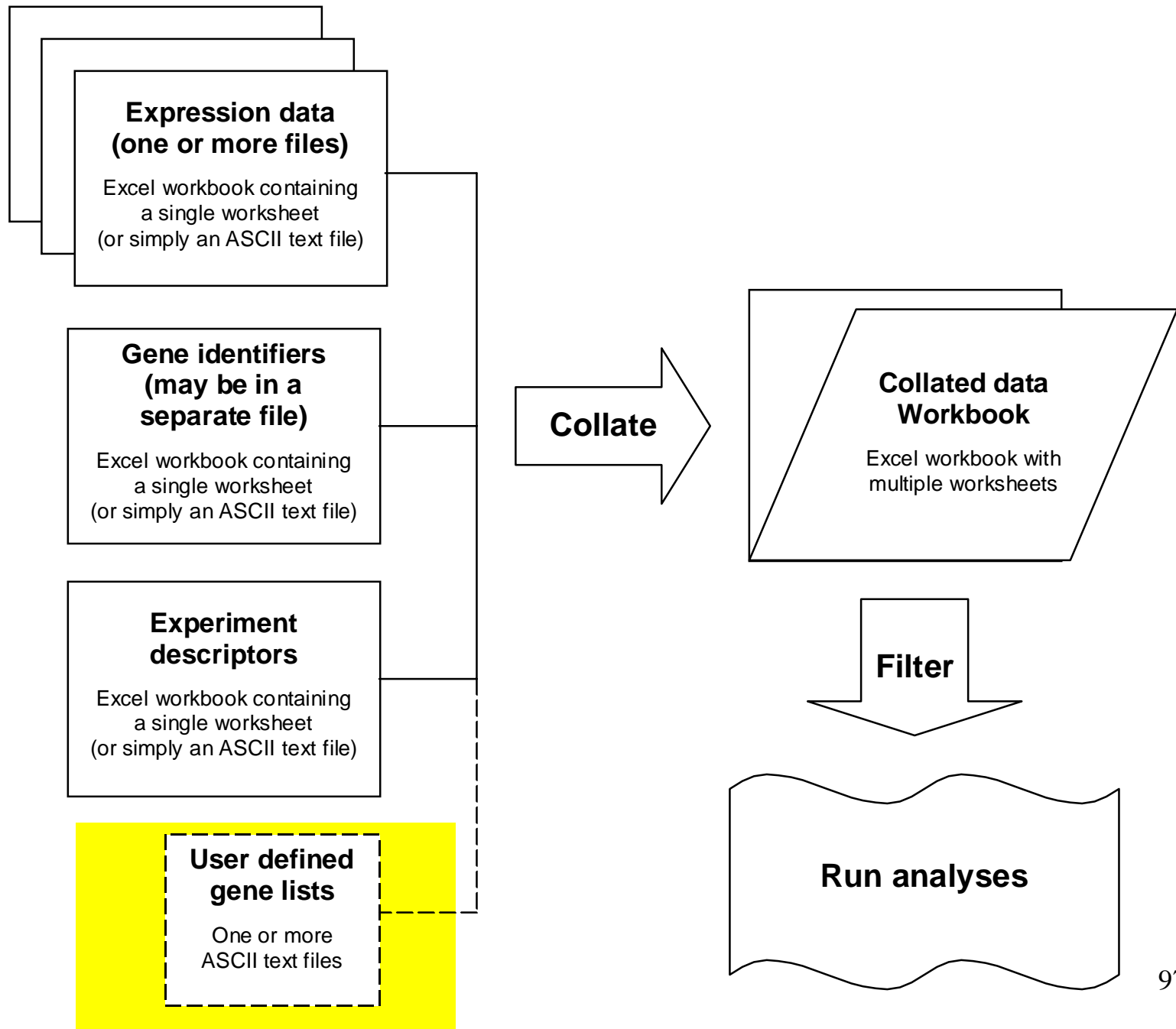
- An experiment descriptors file describes the samples used for each array, and is mandatory.
- For multi-chip sets, use one line per sample, not per array.
- After the header row, each row in this file represents one array or sample, and each column represents one descriptor variable.
- First column 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



| | 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 | | | | | | |



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: CGAP, and user-defined
- CGAP (Cancer Genome Anatomy Project) genelists are pre-loaded with BRB-ArrayTools.
- 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.)

Gene lists

Cancer Genome Anatomy Project

CGI: Angiogenesis - Microsoft Internet Explorer

File Edit View Favorites Tool Links

Address C:\Program Files\ArrayTools\Genelists\CGAP\angiogenesis.html Go

Angiogenesis

- This collection curated by Elise Kohn (ek1b@nih.gov)

| 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 | Angiopietin 1 | D13628, U83508 | AF004327 | |
| ANGPT2 | Angiopietin 2 | AF004327 | | |
| ANGPT3 | Angiopietin 3 | AF107253 | | |
| ANGPT4 | Angiopietin 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 |

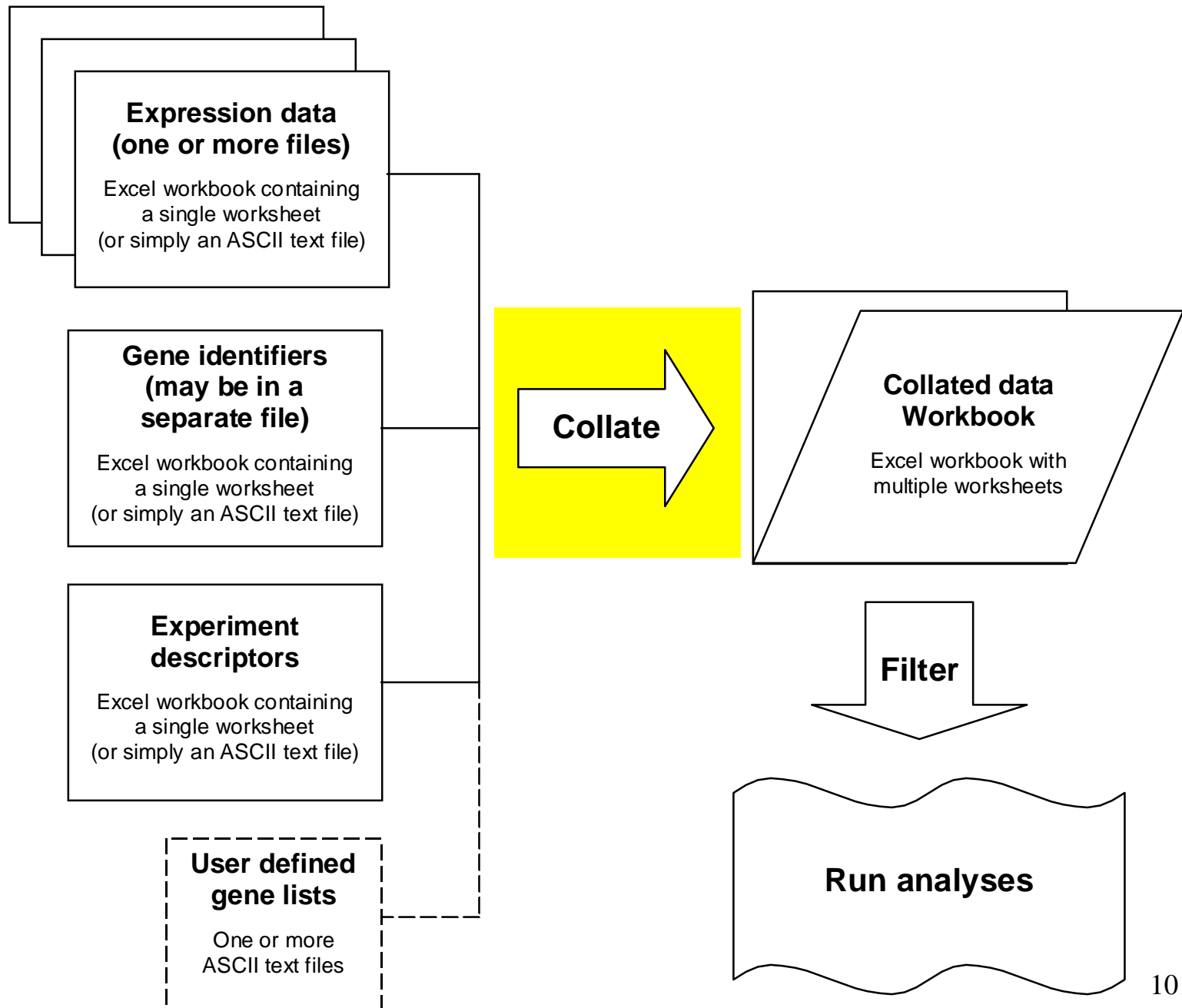
My Computer

Gene lists

User-defined text files

| Perou's- Intrinsic- Breast-Cancer-Genes - Notepad | | | | |
|---------------------------------------------------|----------|------------|-------------|------|
| File | Edit | Format | View | Help |
| Clone | GB acc | UG cluster | Gene symbol | |
| L031045 | AA609880 | HS.1176 | SLC4A3 | |
| L031076 | AA610066 | HS.98428 | HOXB6 | |
| L032796 | AA628430 | HS.425311 | LSM1 | |
| L08422 | T77847 | HS.515951 | SH3YL1 | |
| L08422 | T77926 | HS.515951 | SH3YL1 | |
| L08658 | T72613 | HS.159799 | THRAP2 | |
| L08658 | T72683 | HS.159799 | THRAP2 | |
| L09153 | T81091 | HS.162121 | COPA | |
| L09153 | T81140 | HS.162121 | COPA | |
| L10281 | T71551 | HS.520383 | STX7 | |
| L10281 | T81999 | HS.520383 | STX7 | |
| L20881 | T96082 | HS.99528 | RAB31 | |
| L20881 | T96083 | HS.99528 | RAB31 | |
| L21551 | T97710 | HS.519035 | LAD1 | |
| L21551 | T97813 | HS.519035 | LAD1 | |
| L23614 | R00846 | HS.534072 | C20orf55 | |
| L23614 | R01499 | HS.534072 | C20orf55 | |
| L23980 | R01637 | HS.475963 | CTD SPL | |
| L23980 | R01638 | HS.475963 | CTD SPL | |
| L24781 | R01118 | HS.71465 | SQLE | |
| L28506 | R10154 | HS.513926 | SENP3 | |
| L28506 | R10564 | HS.513926 | SENP3 | |
| L28738 | R09980 | HS.446354 | TCEA3 | |
| L28738 | R16726 | HS.446354 | TCEA3 | |
| L32012 | R24894 | HS.443837 | NPEPPS | |
| L32012 | R32450 | HS.443837 | NPEPPS | |
| L32165 | R23619 | HS.34492 | C10orf32 | |
| L32165 | R26172 | HS.34492 | C10orf32 | |
| L33114 | R26141 | HS.19545 | FZD4 | |
| L33114 | R26355 | HS.19545 | FZD4 | |
| L35118 | R31441 | HS.524134 | GATA3 | |
| L35118 | R31442 | HS.524134 | GATA3 | |
| L35221 | R32848 | HS.2962 | S100P | |
| L35221 | R32952 | HS.2962 | S100P | |
| L35431 | R33004 | HS.547317 | SVEP1 | |
| L35431 | R33005 | HS.547317 | SVEP1 | |
| L36235 | R33642 | HS.523836 | GSTP1 | |
| L36235 | R33755 | HS.523836 | GSTP1 | |
| L38775 | R63543 | HS.448588 | NGFRAP1 | |
| L38775 | R63597 | HS.448588 | NGFRAP1 | |
| L38936 | R62817 | HS.253903 | STOM | |
| L38936 | R62868 | HS.253903 | STOM | |
| L38991 | R62603 | HS.233240 | COL6A3 | |
| L38991 | R62651 | HS.233240 | COL6A3 | |
| L40100 | R65792 | HS.486410 | ECHDC1 | |
| L40100 | R65887 | HS.486410 | ECHDC1 | |
| L40574 | R66139 | HS.410554 | CX3CL1 | |
| L40574 | R66139 | HS.531668 | CX3CL1 | |

| HG-U95_Housekeeping - Notepad | | | | |
|-------------------------------|------|--------|------|------|
| File | Edit | Format | View | Help |
| Probe set | | | | |
| 34864_at | | | | |
| 39782_at | | | | |
| 39415_at | | | | |
| 36928_at | | | | |
| 39047_at | | | | |
| 38483_at | | | | |
| 41833_at | | | | |
| 41224_at | | | | |
| 38016_at | | | | |
| 35753_at | | | | |
| 31385_at | | | | |
| 40281_at | | | | |
| 905_at | | | | |
| 39866_at | | | | |
| 39027_at | | | | |
| 39336_at | | | | |
| 32518_at | | | | |
| 1315_at | | | | |
| 1009_at | | | | |
| 39811_at | | | | |
| 41785_at | | | | |
| 32437_at | | | | |
| 31907_at | | | | |
| 39360_at | | | | |
| 36587_at | | | | |
| 35835_at | | | | |
| 1310_at | | | | |
| 39184_at | | | | |
| 31573_at | | | | |
| 41295_at | | | | |
| 36972_at | | | | |
| 33656_at | | | | |
| 1653_at | | | | |
| 36167_at | | | | |
| 38817_at | | | | |
| 31952_at | | | | |



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 data import wizard can be used to guide you through the specification of the data components
- mAdb data archives: Please see separate handout for specific instructions on downloading the formatted archive from mAdb.
- GenePix: Specify the folder containing the .GPR files and in addition you can import gene identifiers from the .GAL or .GPR file
- Affymetrix data: Automatically imports data by searching for “Probe Set Name”, “Signal” (or “Avg Diff”), and “Detection” (or “Abs_Call”) column header labels. For complete details please refer to the User’s Manual.

Affymetrix CEL file importation

- For importing Affymetrix CEL files, go to the following menu items: **Collate data → Special format: Affymetrix GeneChips → Probe-level data (.CEL files)**
- You will need to browse for a data folder containing the .CEL files, and provide an Experiment Descriptors file. Gene identifiers will be imported automatically from the BRB server.
- This utility currently uses the RMA/GC-RMA functions included in the 'affy'/'gcrma' package of BioConductor. Future versions of BRB-ArrayTools will include other methods for computing expression summaries.
- **WARNING:** The RMA functions require large memory capacity and the run-time may be slow (up to several hours for more than 100 arrays).

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>

Part VI:

Independent practice
(if time permits)

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

When sending files to accompany bug reports, please send attachments SEPARATELY from the text of your bug report. This is to ensure that we receive the text of your bug report even if the attachments are blocked either on the sender's end or receiver's end. Also, change or remove all .zip file extensions before sending files.

BRB-ArrayTools ListServ

To participate in ListServ, send email to

listserv@list.nih.gov

with the following in the MESSAGE BODY:

subscribe BRB-ArrayTools-L yourname

Please refrain from sending attachments with your ListServ messages. If a particular ListServ member requests to see a file, please send attachments individually to that member.

Once subscribed, you can always unsubscribe or set your subscription to DIGEST mode later.

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!!