Microarray Data Analysis Using BRB-ArrayTools Version 4.2.0 –Beta_2

Supriya Menezes

BRB- ArrayTools Development Team

May 17h 2011

Agenda

- What is BRB-ArrayTools?

I. Installing BRB-ArrayTools and its required components.
II. Creating a collated project workbook.
III. Data filtering and normalization options
IV. Break.
V. Graphics
VI. Class Comparison
VII. Gene Set comparison
VIII. Clustering
IX. MDS
X. Class Prediction
XI. Plug-ins
XII. Tutorial.
XIII. 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 2.12.0. 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/MAS5.0 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. As well as on an Apple macbook pro machine with Windows OS installed with Apple's bootcamp software
- Currently compatible with MS Vista/ Windows 7 and Excel 2007/ 2010 .
- Also works on a 64- bit machine with Windows OS.
- 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.

New to ArrayToolsv4.1

- Affy ST array importer.
- Enhanced visualizations and interactive plots.
- Enhanced the Heatmap in clustering.
- New plug-ins: Adaboost and Lassoed PC.
- A new gene filtering to handle redundant probe sets that correspond to the same gene.
- Utility: To obtain drug information based on a gene list.
- Ability to import custom expression arrays and annotations by using the gene identifiers.

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-ArrayToolsv4.1.0
- Additionally, v4.2 beta release.

Full Installer

- Also available is an option to download a FULL installer. This file is a bundle of all the necessary components like Rv2.12, statconnDCOM and java are included along with ArrayTools and CGHTools.

Installing BRB-ArrayTools Pre-download

Downloading BRB-ArrayTools

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

NATIONAL CANCER INSTITUTE

Biometric Research Branch

Division of Cancer Treatment and Diagnosis

BRB ArrayTools

Developed by: Richard Simon & BRB-ArrayTools Development Team

The software is free for non-commercial use. Commercial users should contact Michael Shmilovich at shmilovichm@od.nih.gov or (301)435-5019.

If you do not have a password, please go to our GUESTBOOK and make an application.

If you forgot your password, please enter the email address you used for registration. We will send you the password.

E-mail: [ ] Submit

BRB-ArrayTools Download

Please enter your password in BOTH username and password fields when prompted.

Download Standard Version 3.8.1

Download Commercial Version 3.8.1

Download 60-day Trial Version 3.8.1

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

*Instructions for Vista Users to take Full Control of the ArrayTools installation folder.

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

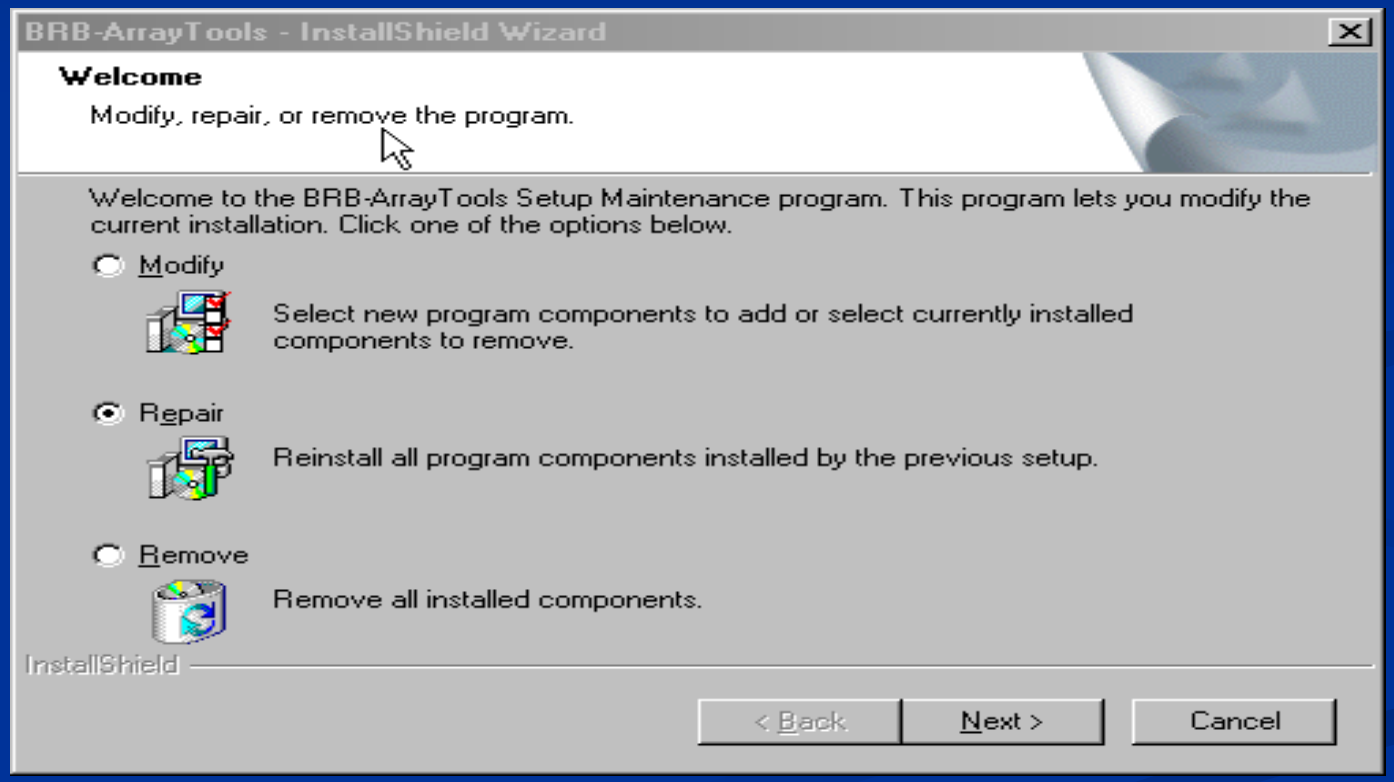
Installing BRB-ArrayTools

- On your desktop look for the folder called "BRB-ArrayTools-Class".
- Run the file called "ArrayTools_v4_2_0_Beta_2_Full.exe".

Installing BRB-ArrayTools

Installing BRB-ArrayTools

- Select "**Repair**" option and click "**Next**" button.



Installing BRB-ArrayTools

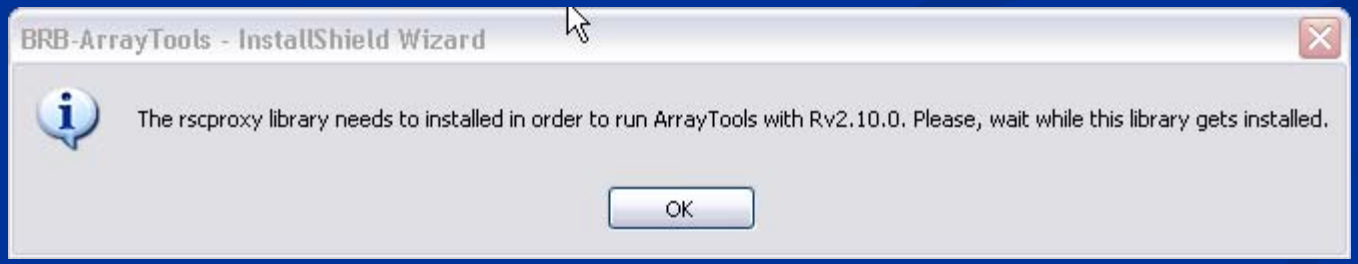
- Select "Yes" to the question about Administrator privileges on the computer.

Installing BRB-ArrayTools

- Click “OK” to install R, RCOM and Java.

Installing BRB-ArrayTools

- Proceed to install Rv2.12.0 using all the default options.
- Complete the set-up of R.
- Click "OK" to install the rscproxy package.

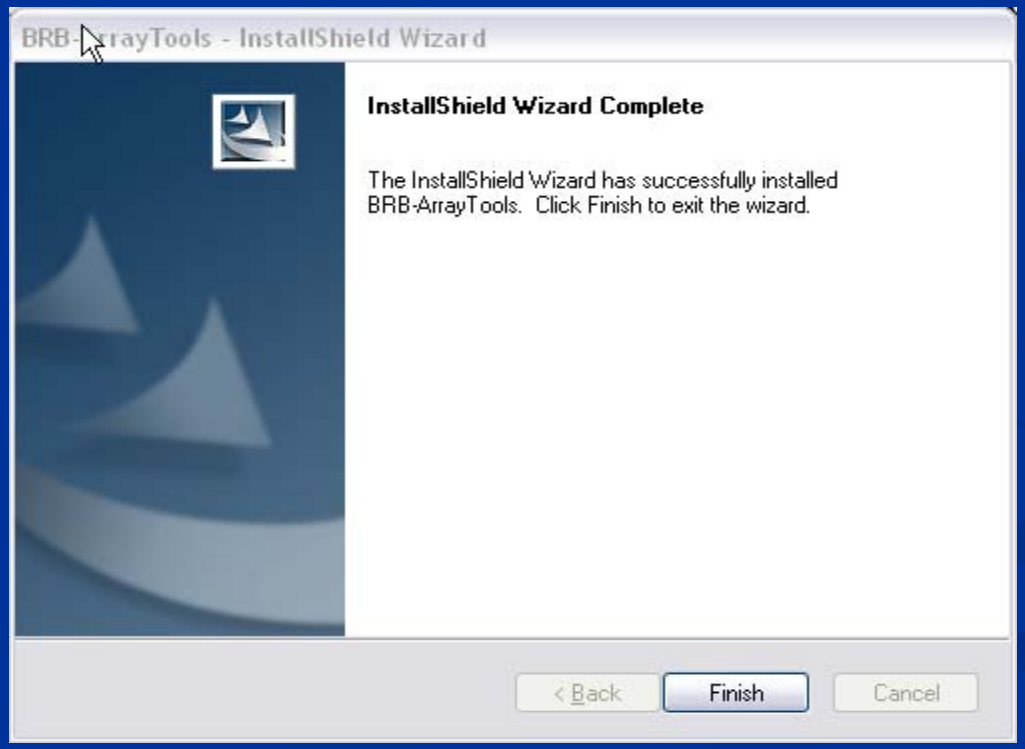


Installing BRB-ArrayTools

- Proceed to install RCOM and Java using the default options.
- Install CGHTools.

Installing BRB-ArrayTools

- Installer will install BRB-ArrayToolsv4.1 and you will get the message below. Click on the "Finish" button.



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- 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, RServer 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, all the three Add-Ins should be listed there namely: ArrayTools, CGHTools and RServer add-ins.

[Hands-on instructions]

[Getting started]

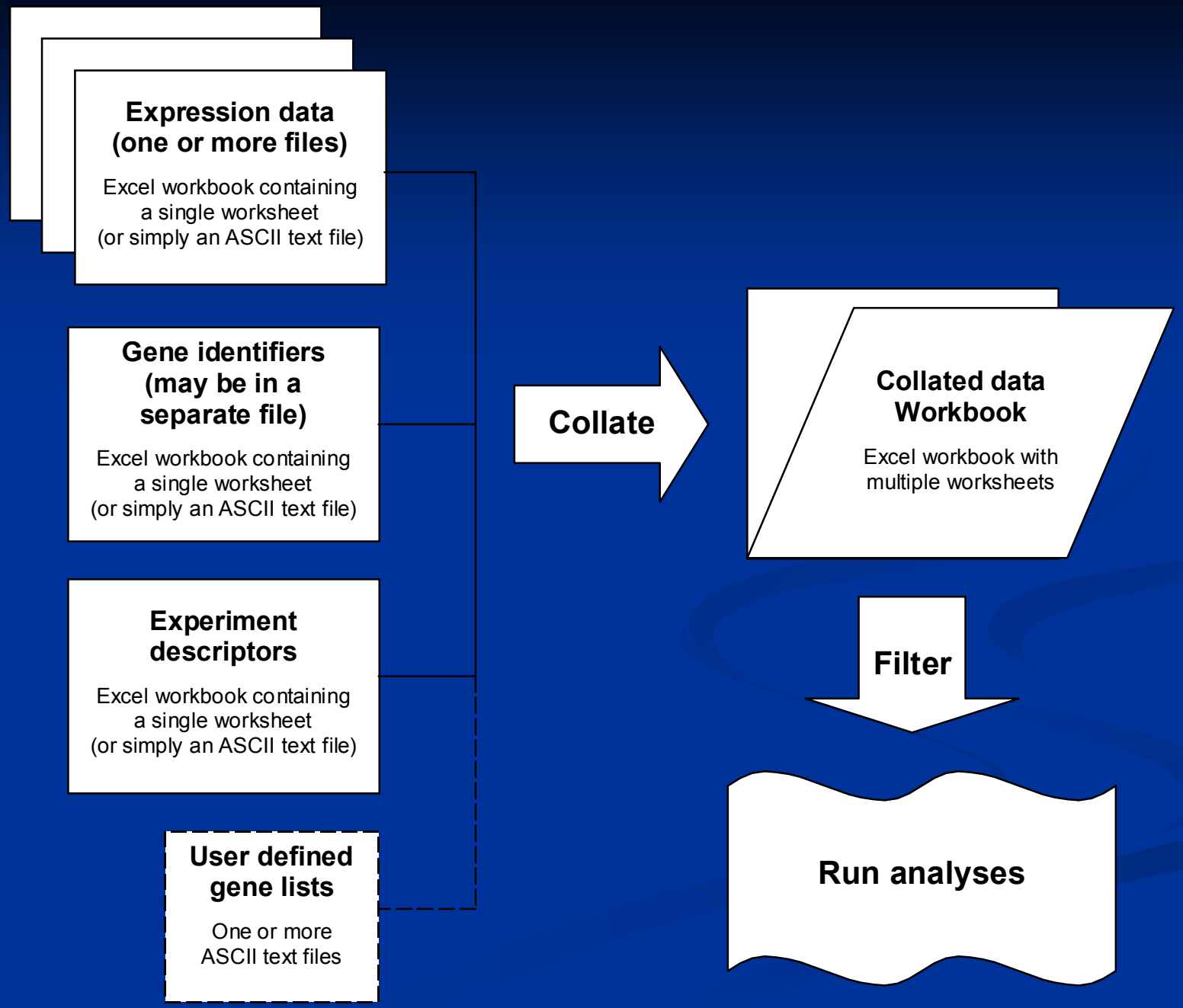
1. Open **Excel**.
2. Click on **Tools →Add-ins**, and see that **BRB-ArrayTools** is loaded as an add-in.
3. When BRB-ArrayTools is loaded as an add-in, you will find an **ArrayTools** menu. This is the interface for all BRB-ArrayTools functions.
4. Click on **ArrayTools → Getting started**.
5. Here 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".

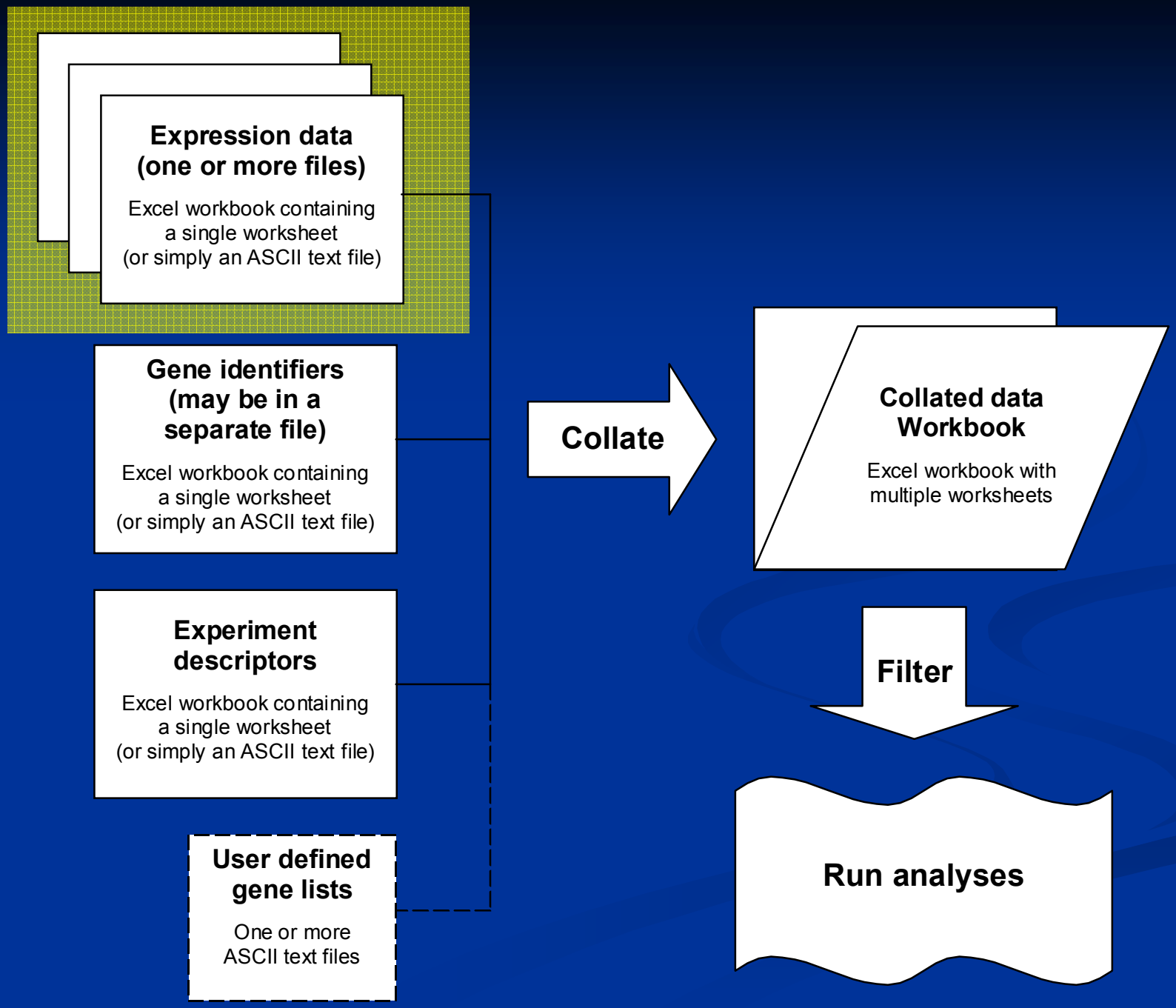
VISTA Users – Full Control to ArrayTools folder (optional)

- 1: Open the windows explorer (Windows key + E)
- 2: Go to the “C:\Program files”, right click the mouse on the “ArrayTools” folder.
- 3: Pick the “Properties” at the bottom of the menu.
- 4. Select the “Security” tab.
- 5. Click on your “User Name”. In this slide, we assume the user name is “BRB_VISTA”.
- 6. Then click on then “Advanced” button.
- 7: Click on the “Owner” tab.
- 8:It shows the folder owner is Administrators. If you are not Administrator.Talk to your technical support for help.
- 9. Click “Edit” button. You will see the message
- 10. Windows needs your permission to continue.
- Click “continue” button.
- 11: Select your “UserName” and then the “Apply” button.
- 12. You may get the following message. Just ignore it by clicking “OK” button.
- 13. Click “OK” button once more.
- 14. As you can see the folder in the next screenshot, the owner of the folder is changed to your “UserName”.
- 15: Now, click the “OK” button to return to the folder’s Properties. We are still at the “Security” tab of the Properties.
- Click on your “UserName”, and then 'Edit' button.
- 17. Click on “Full Control” and “Allow"
- 18. Now, Click “OK”.
- 19: Now, click on “Apply” and then “OK”

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 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.

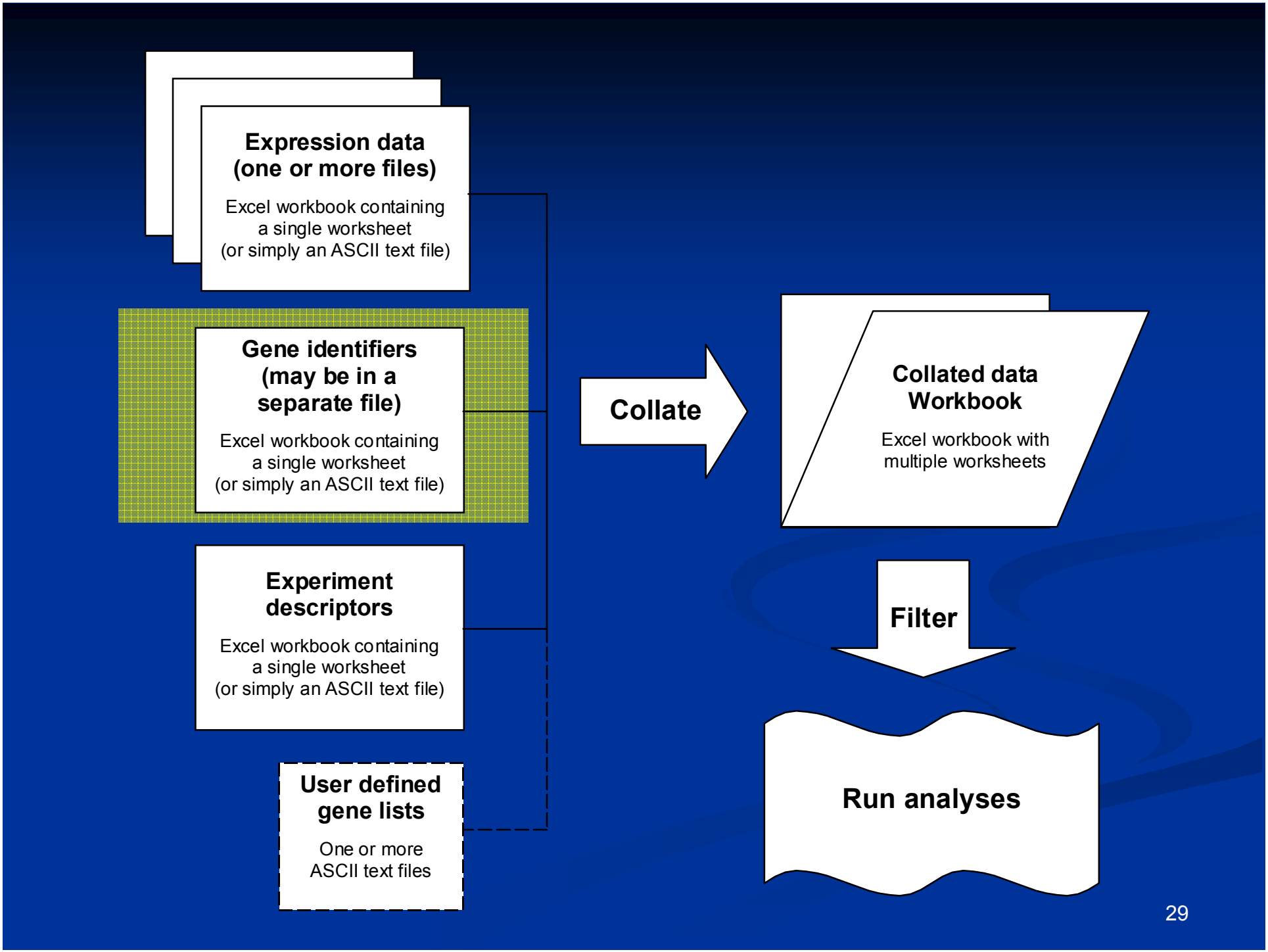
Expression data

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_
2	600001	IMAGE:604856	adhesion selectin B Mm	21363	13268	0	19674	11840	0	11938	4870	
3	600002	IMAGE:619876	adhesion VCAM-1 Mm.1	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 thrombospodir	14204	6937	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	

Horizontally stacked example


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 accessions, and probe set ids.

Gene identifiers

Two examples of a gene identifier file

GeneIds.xls

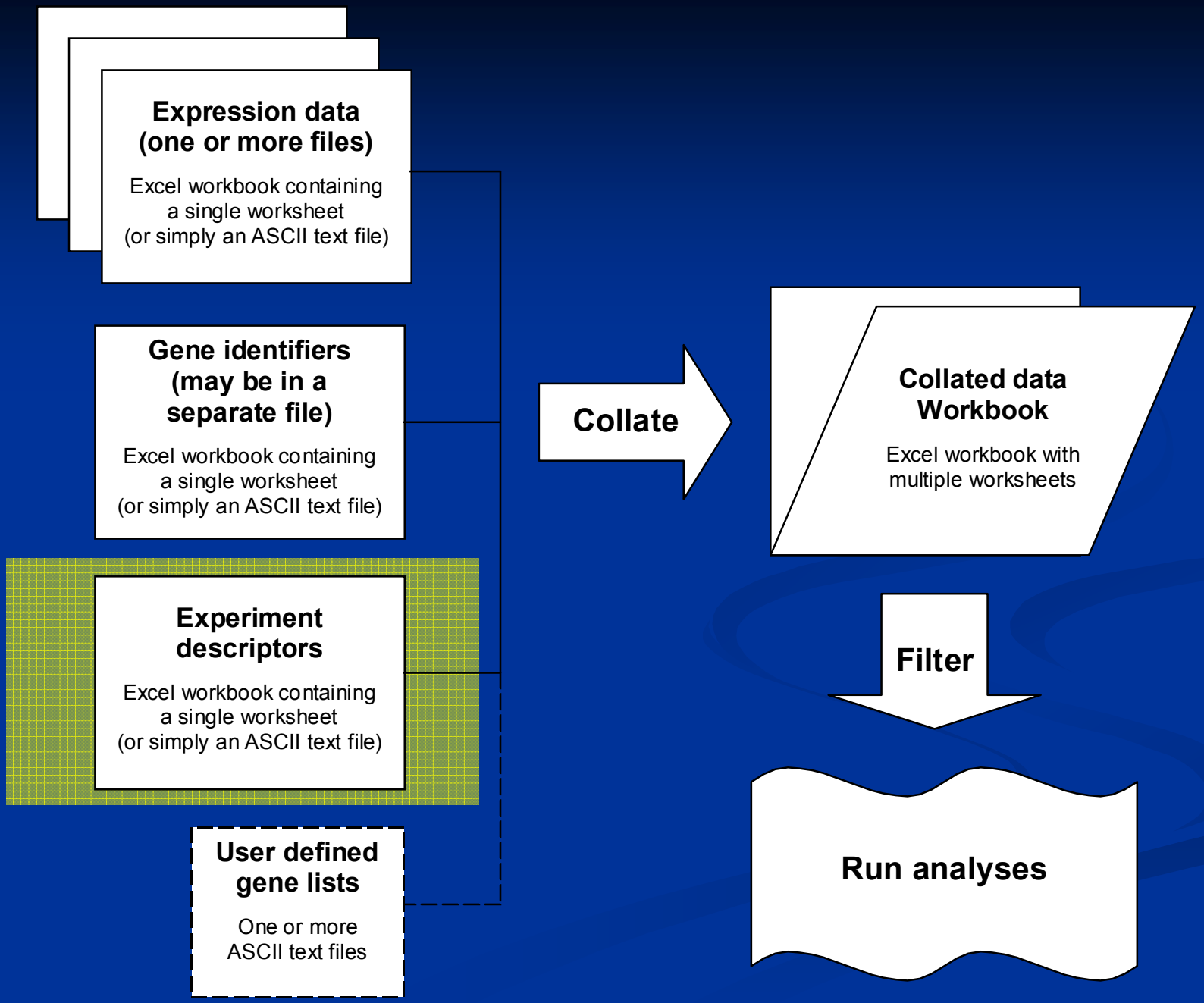
	A	B	C	D	E
1	Spot	Clone	Description	GB acc	
2	49	60204	Homo sapiens C2H2 zinc finger protein pseudogene, mRNA sequence	T39154, T40438	
3	50	60436	RPL3 Ribosomal protein L3 Chr.22	T39295, T40510	
4	51	60218	ESTs	T39165, T40450	
5	52	60209	ESTs	T39163, T40448	
6	53	60664	ESTs	T39448, T40595	
7	54	60932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


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.
- 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.

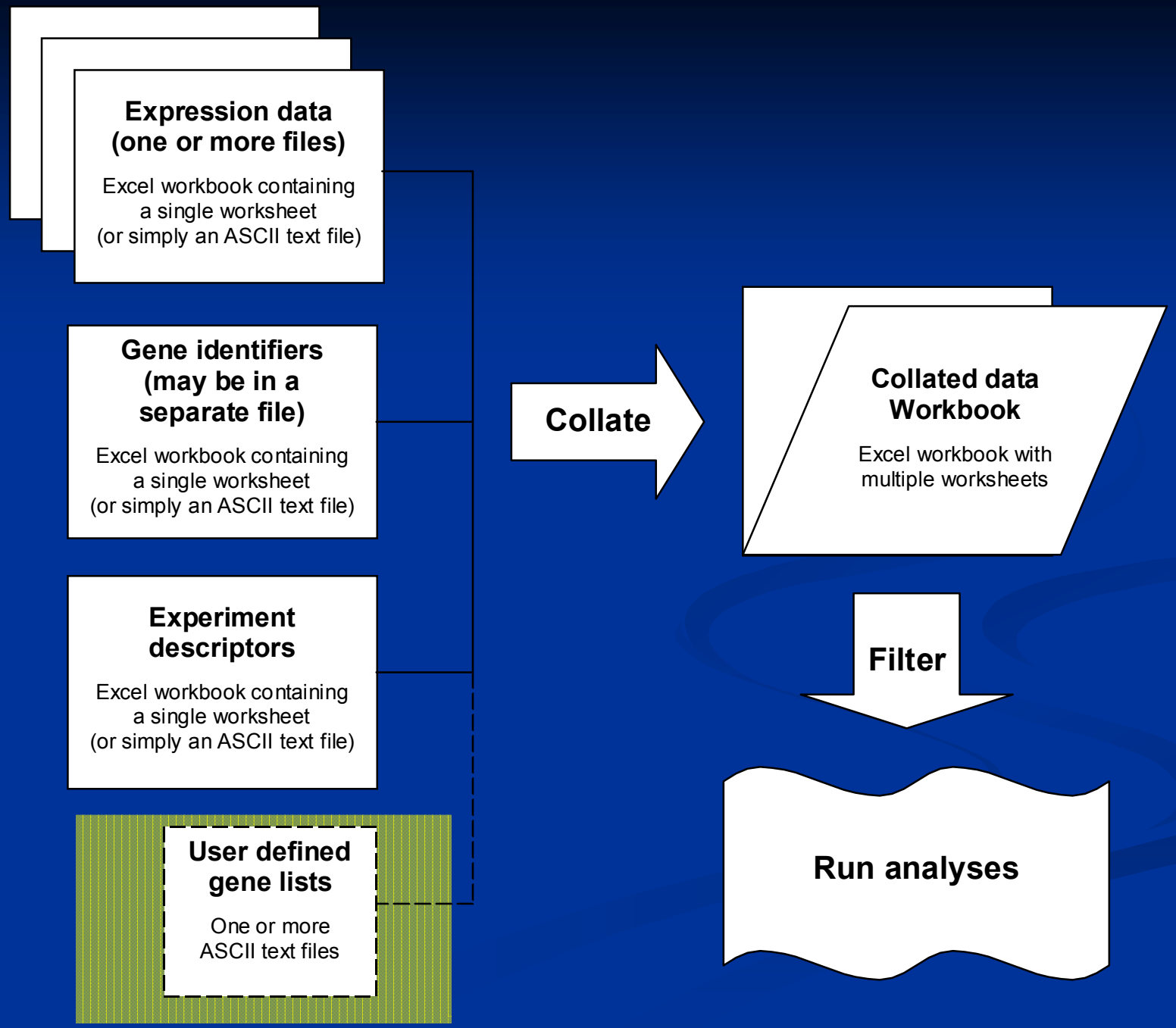
Experiment descriptors

Describes the samples used for each array

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: 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

CGL: 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	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

My Computer

Gene lists

User-defined text files

Perou's- Intrinsic- Breast-Cancer-Genes - Notepad

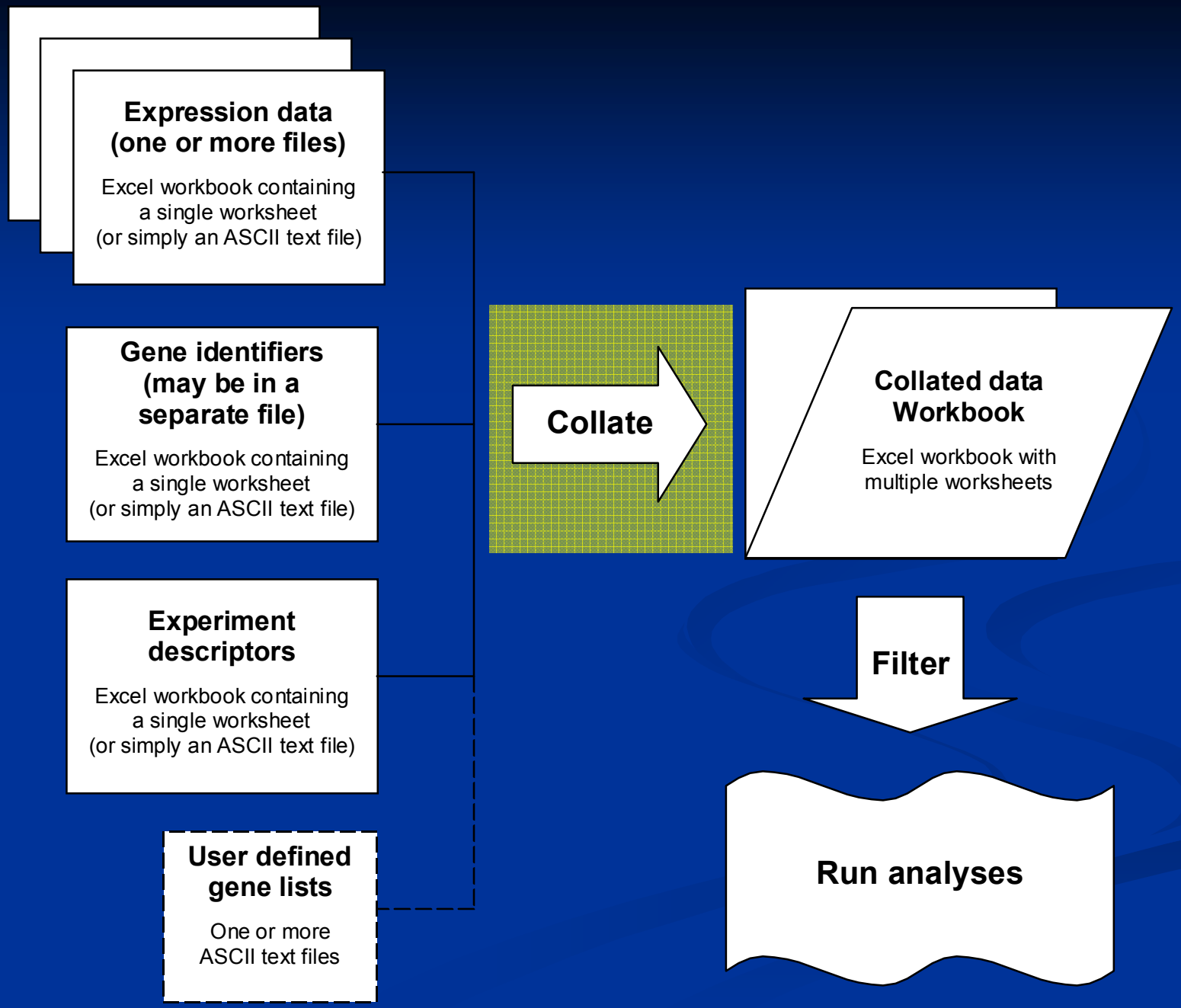
File Edit Format View Help

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	CTDSPL
L23980	R01638	Hs.475963	CTDSPL
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


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 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: (**Data Import Wizard**)
- 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.
- Additionally can compute MAS5.0 summaries from .CEL files.
- For large number of arrays (more than 100), a new method called 'almostRMA' is available that avoids previous memory limitations.

Recently Implemented

- Can automatically import a GDS dataset from the NCBI Gene Expression Omnibus (GEO) database into BRB-ArrayTools.
- Can directly import dual channel Agilent data into BRB-ArrayTools using the data import wizard.
- Ability to import illumina data using the data import wizard with the lumi package.

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:

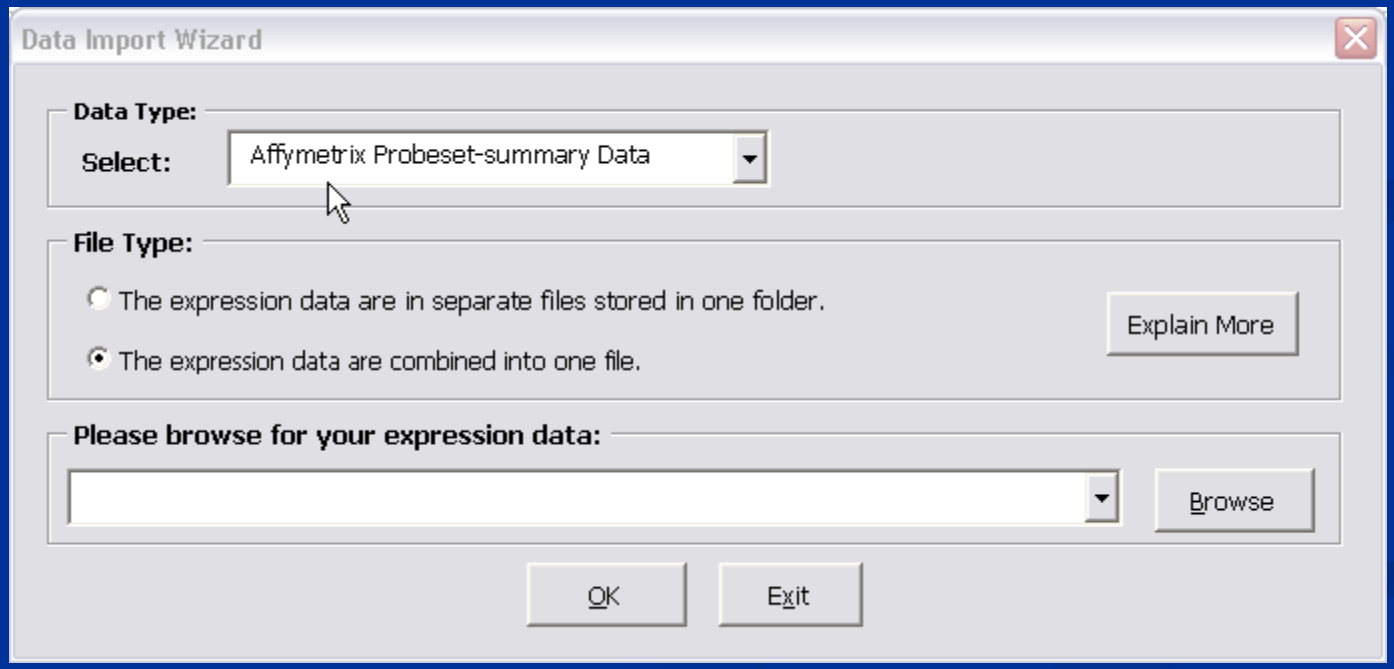
  Dataset_A2_multiple_tumor_samples.txt
  ExpDescrMedulo.xls

- 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]

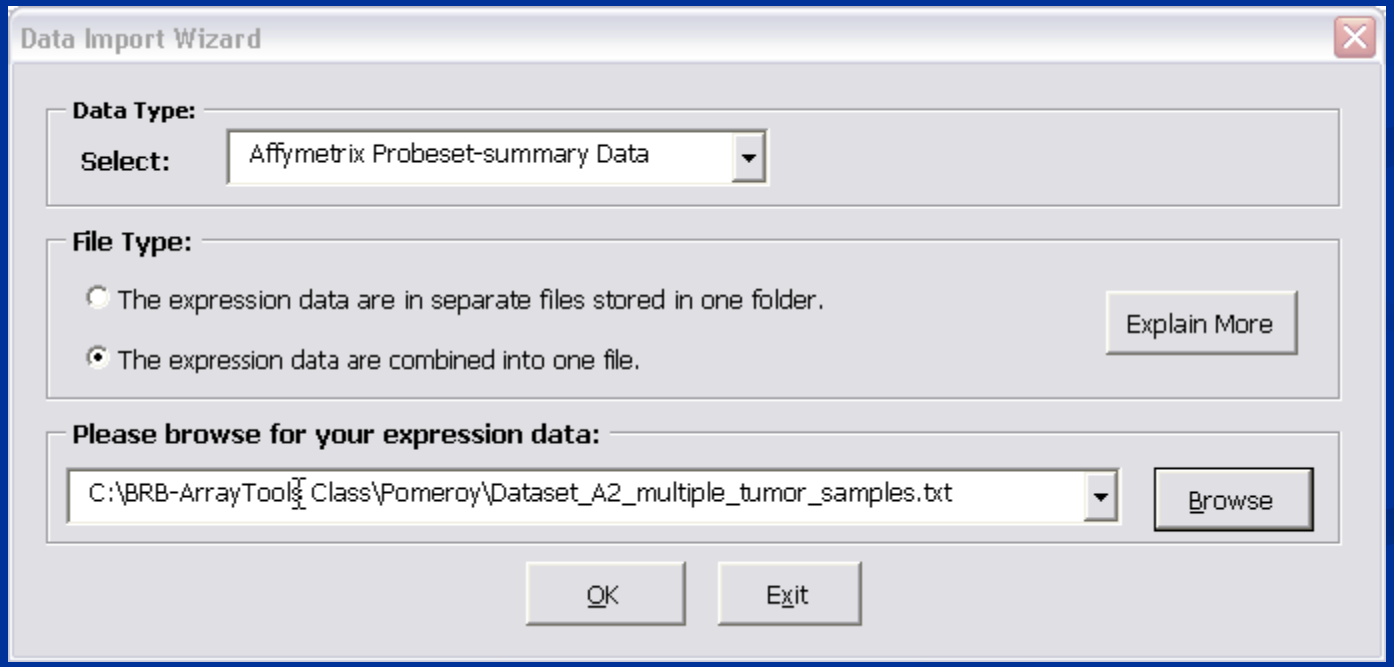
- Click on **ArrayTools →Getting started → Data Import Wizard**
- Select the option from the pull down menu- **"Affymetrix probeset-summary data".**
- **Choose the option that the expression data is combined into one file.**



[Hands-on instructions]

[Importing Pomeroy Data set]

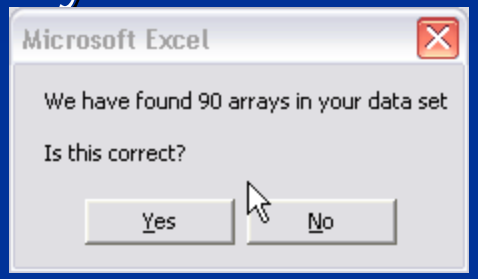
•**Browse** for the following file which is also in the **Pomeroy** folder inside the **BRB-ArrayTools Class** folder which is on the **Desktop**: `Dataset_A2_multiple_tumor_samples.txt` and then click **OK.**



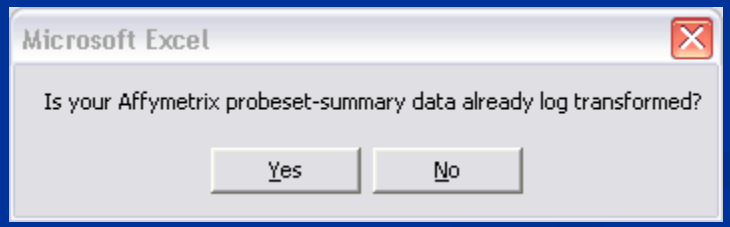
[Hands-on instructions]

[Importing Pomeroy Data set]

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



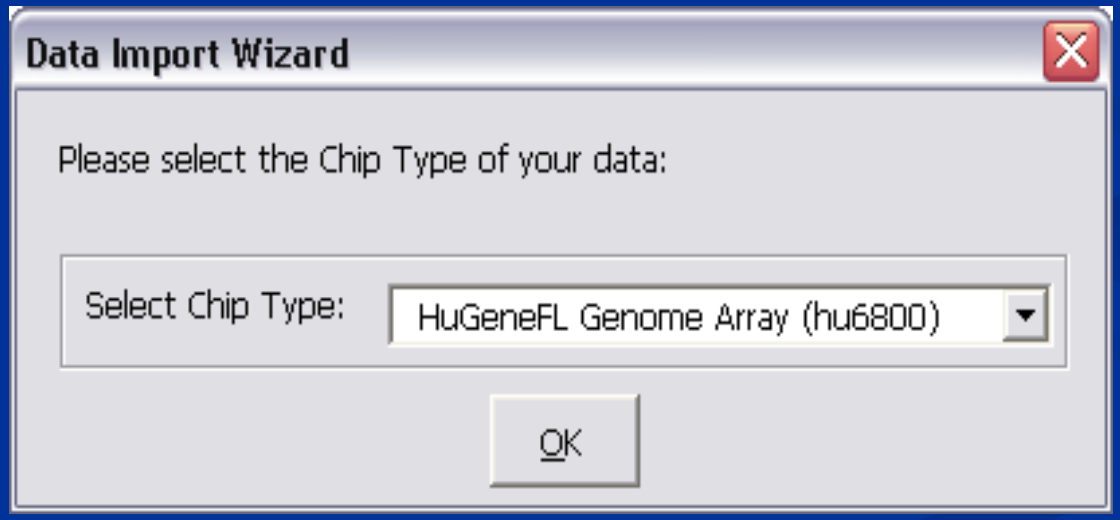
- Click "No" to the question about log transformation.



[Hands-on instructions]

[Importing Pomeroy Data set]

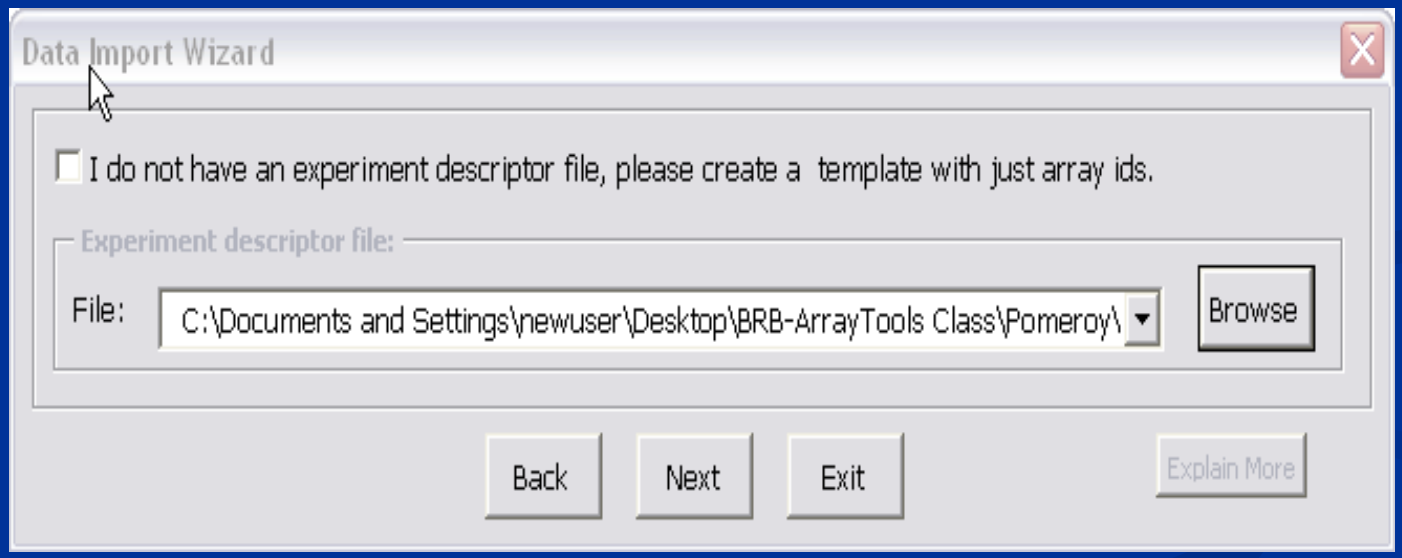
- Select the chip type as "HuGeneFL Genome Array"



[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".



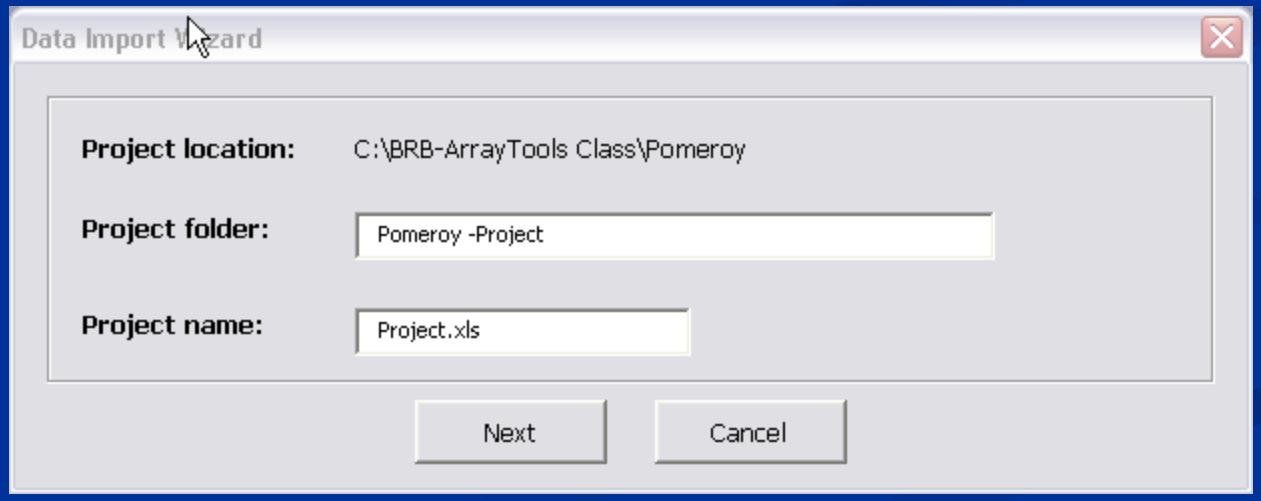
[Hands-on instructions]

[Importing Pomeroy Data set]

- Keep the defaults for Filtering.

Save the Project in the folder "Pomeroy-Project".

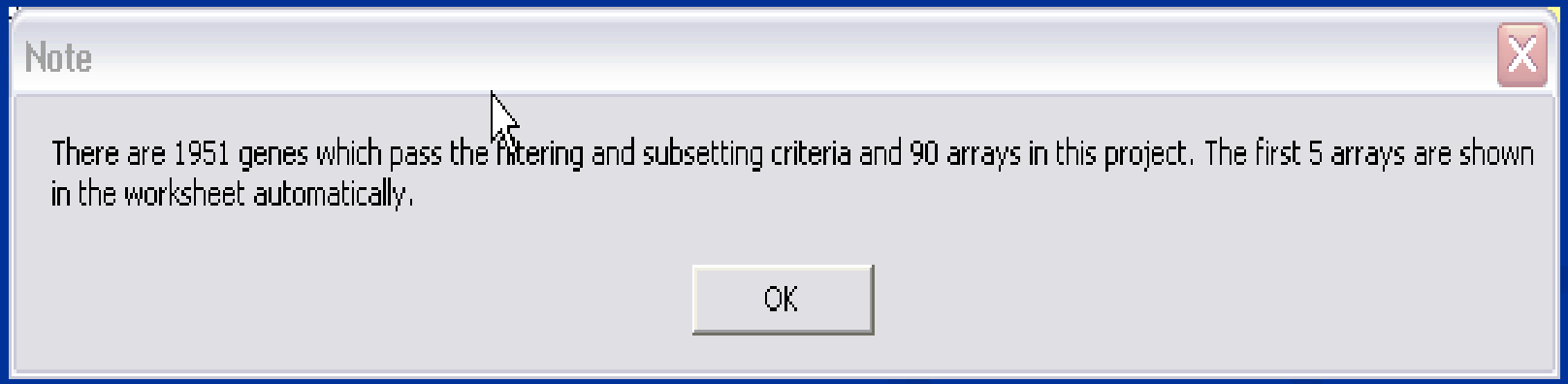
- The progress bar will indicate that the project is collating.



[Hands-on instructions]

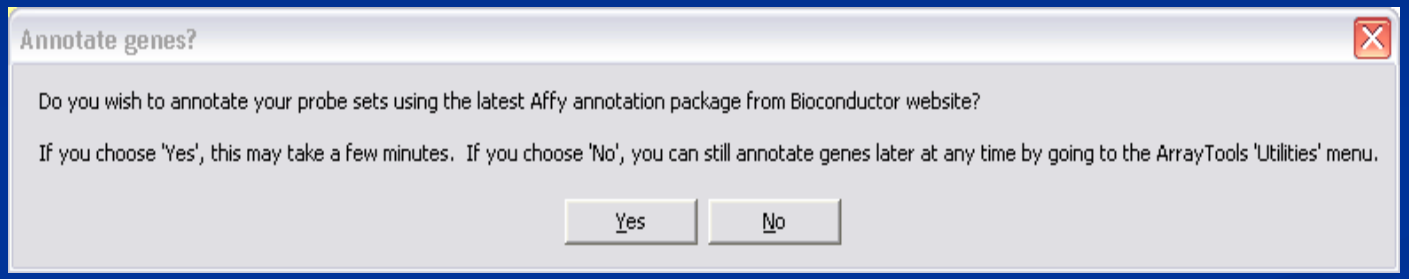
[Importing Pomeroy Data set]

- Click "OK"



[Hands-on instructions]

[Importing Pomeroy Data set]



Click **Yes** to annotate the project.

Collated project workbook

Overview

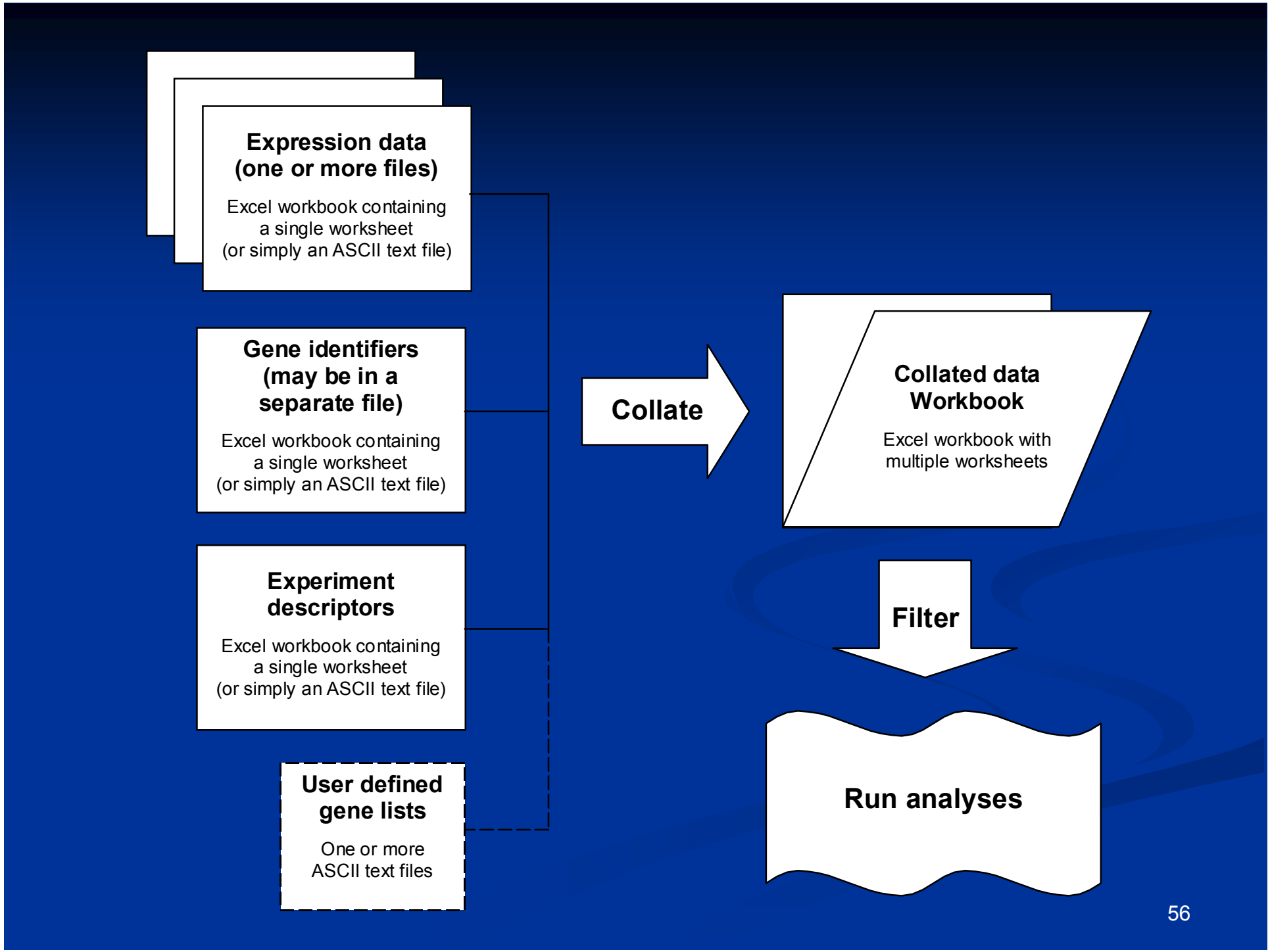
- The **collated project workbook** is the primary data object on which future analyses are run
- The collated project workbook is located inside the **project folder**, which by default is located inside the folder where the original input data is located.
- The project folder may also contain some 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.

Microsoft Excel - Project.xls

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

A1 = Array

	A	B	C	D	E	F	G	H	I	J	K	L
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	T2MO	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

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

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

Arial 10

A1 = Probe set

	A	B	C	D	E	F
1	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 exchar	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 interactir	NPIP /// LOC283970 /// LOC339047 /// LOC3		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 binc	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	

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

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

Arial 10

B2 = 0

Click to display the data

1	Probe set	Missing	P-Value	Rank	Variance	Num 1.5-Fold	Absent (Affy)	Filter
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

Arial 10

A1 = Probe set (Double-click)

	A	B	C	D	E	F	G	H	I	J
1	Probe set (Double	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	chromosor	AB000460	Hs.125348	C4orf8	8603	4	Chr:4p16.3		TRUE
13	AB000464	chromosor	AB000464		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 matura	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	TRUE
55	AC002045	nuclear po	AC002045	Hs.446747	NPIP /// LC	283970 ///	16	Chr:16p13-	biological	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	TRUE
106	AF008937	syntaxin 1	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	AFFX-HSA	actin, beta	X00351	Hs.520640	ACTB	60	7	Chr:7p15-p	cellular co	TRUE
140	AFFX-HSA	actin, beta	X00351	Hs.520640	ACTB	60	7	Chr:7p15-p	cellular co	TRUE
141	AFFX-HSA	actin, beta	X00351	Hs.520640	ACTB	60	7	Chr:7p15-p	cellular co	TRUE

Experiment descriptors | Gene annotations | Filtered log intensity | Gene identifiers

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 50th 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.

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 between groups of arrays

This procedure finds genes differentially expressed among classes of samples. The classes are pre-defined based on columns of the experiment descriptor file. Each array should represent one sample, either as a single-label experiment or as a dual-label experiment using a common reference. For non-reference designs, consider using the tool for class comparison between red and green samples.

Experimental design:

Column defining classes:

Unpaired samples:

Block by:

Average over replicates of:

Paired samples:

Pair samples by:

Find gene lists determined by:

Significance threshold of univariate tests: 0.001

Restriction on proportion of false discoveries:

Maximum proportion of false discoveries: 0.1

Confidence level (between 0 and 100%): 80

Restriction on number of false discoveries:

Maximum number of false discoveries: 10

Confidence level (between 0 and 100%): 80

Variance model:

Use random variance model for univariate tests.

NOTE: This analysis is currently set to run on all genes passing the filter.

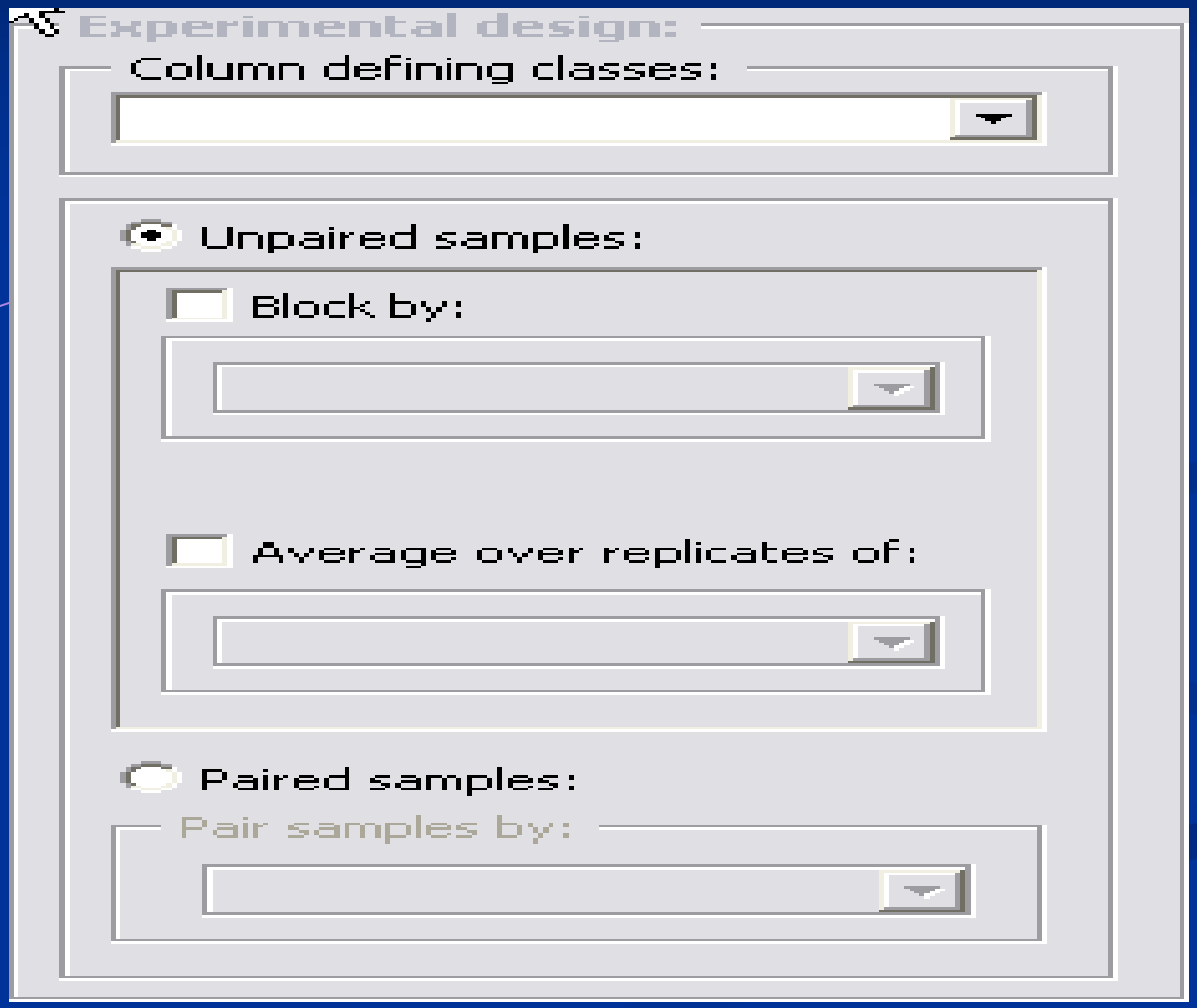
Select gene subsets

OK | Cancel | Options | Reset | Help

Class Comparison
Experimental design

Step1

Step2



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

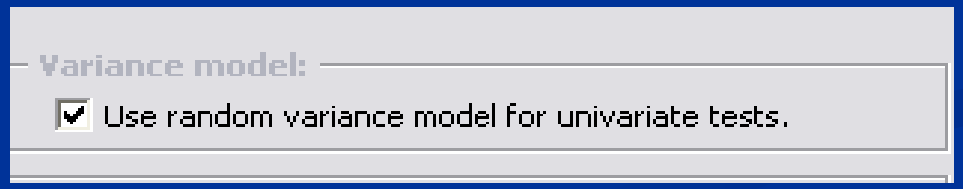
Average over replicates

- 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:

Find gene lists determined by:

- (•) Significance threshold of univariate tests: 0.001

- ( ) Restriction on proportion of false discoveries:

  Maximum proportion of false discoveries: 0.1

  Confidence level (between 0 and 100%): 80

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

Find gene lists determined by:

- ○ Significance threshold of univariate tests: 0.001
- ● Restriction on proportion of false discoveries:

  Maximum proportion of false discoveries: 0.1

  Confidence level (between 0 and 100%): 80

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 **90th 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.
- New to v3.8, the ability to handle multiple probe sets that correspond to the same gene either using the average intensity (single channel data only) or inter quartile range.

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'

Goeman's Global test

- Tests the null hypothesis that no genes within a geneset are differentially expressed.
- A gene set is said to be significant if the corresponding parametric global p-value is less than the threshold value selected by the user.

[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.

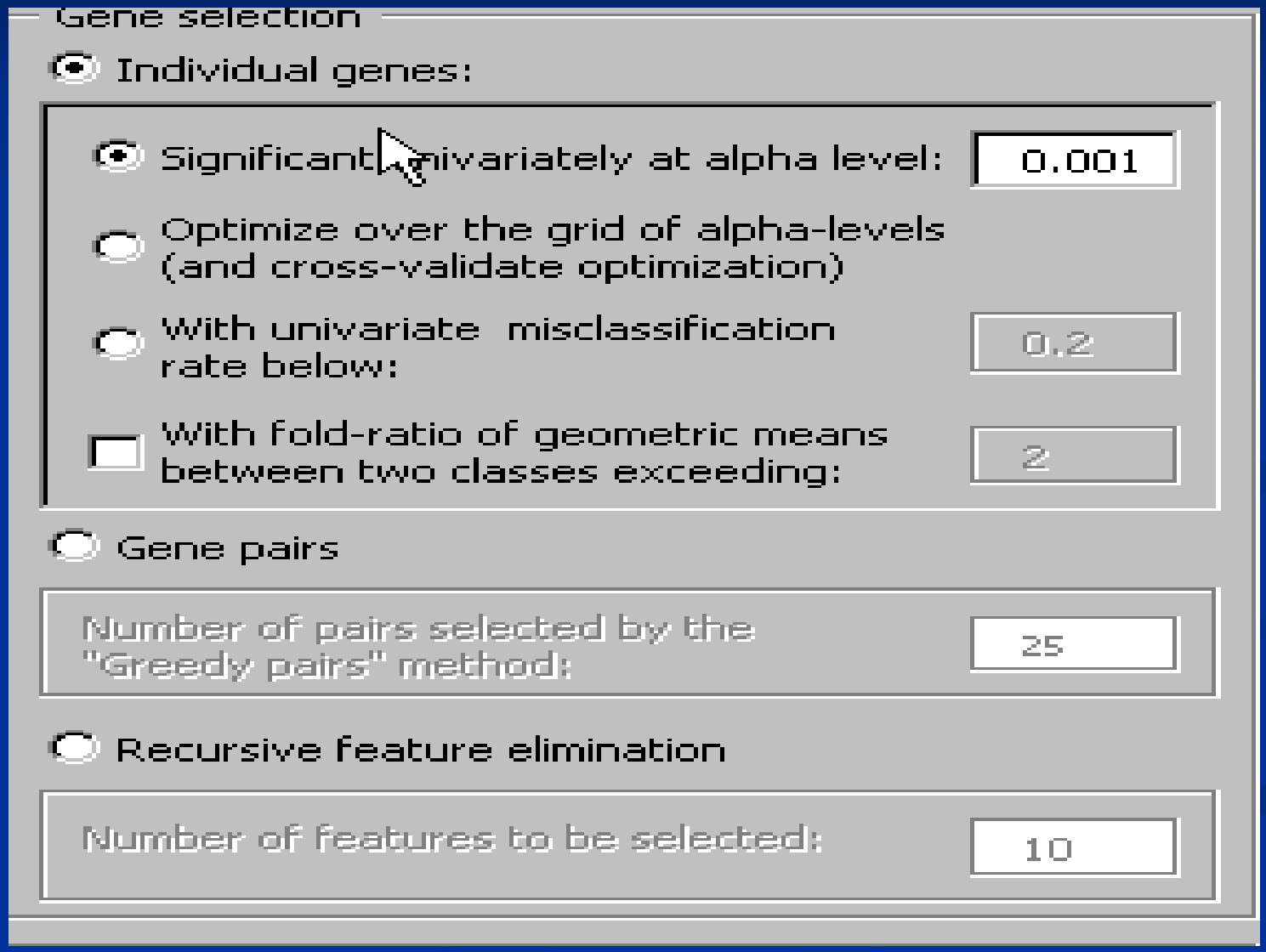
- C3. Evaluating the Classifier

  - Cross-validation

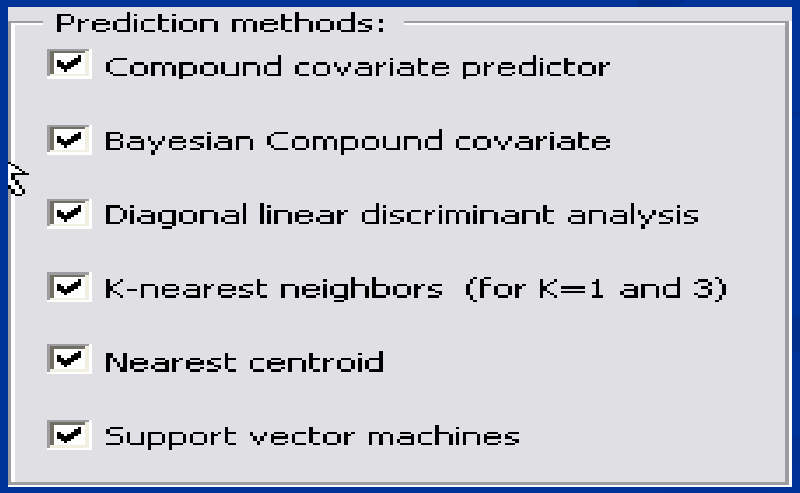
C1. Gene Selection Criteria

- 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
- New to v3.6, is the Recursive feature elimination method.

Gene Selection Criteria

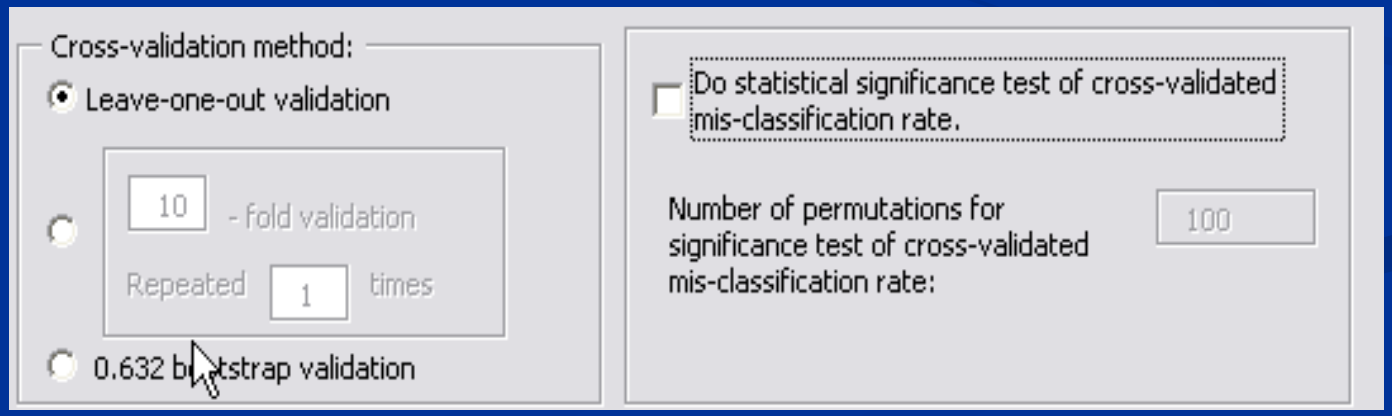


C2. Class prediction Methods

- 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:

$$\Sigma_i t_i x_i \quad \begin{cases} \text{where } t_i = \text{t-statistic}, \; x_i = \text{log-ratio}, \\ \text{and sum is taken over all significant genes} \end{cases}$$

- 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 Branch 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

  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-ArrayToools 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!!

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.