Microarray Data Analysis

A step by step analysis
using BRB-Array Tools

EXAMINATION OF DIFFERENTIAL GENE EXPRESSION (1)

- Objective: *to find genes whose expression is changed before and after chemotherapy*.
- Experiment: Biopsies from breast cancer patients, before and after a 16 week course of doxorubicin chemotherapy.
- Reference: Korn et. al. *Identifying Pre-Post Chemotherapy Differences in Gene Expression in Breast Tumors.*

Study design

- RNA samples from 20 breast cancer patients tumors.
- cDNA microarrays.
- Reference design: each tumor sample was compared with pooled mRNA from 11 cell lines.
- Paired data: two samples (marrays) per patient: one before and one after chemotherapy.
- For the analysis here, a subset of 2998 genes will be used.

Analysis pipeline

1. Load (collate) the data.
2. Filter bad spots &
   Adjust for low intensities.
3. Normalize and check.
4. Do the tests.
5. Interpret results.

1. Load data. Option 1: Collate

- If we are going to work with our own data (.CEL or .gpr files) or with data obtained from a database we must import it into the format used by BRB.
- This can be done following the steps in *Array Tools → Import data → Import wizard*
- In this tutorial we will use an example project that has already been created.

1. Load data. (2): Existing project

- A BRB project workbook with the prepared data is available in the “sample datasets” folder. Its name is "Perou.xls.“
- Load the project and inspect the four worksheets it contains:
  - **Experimental Descriptors.**
  - **Gene Identifiers.**
  - **Gene Annotations.**
  - **Filtered Log Ratios.**

Experimental Descriptors

EXP	PatientID	BeforeAfter
svcc77	10	AF
svcc78	10	BE
svcc86	100	AF
svcc104	100	BE
svcc85	101	AF
svcc84	101	BE
svcc82	102	AF
svcc101	102	BE
svcc65	104	AF
svcc120	104	BE
svcc121	109	AF
svcc105	109	BE
svcc126	112	AF

Samples (one per chip)

Covariates
(Other usual ones Might have been "SEX", "TREATMENT"

Gene Identifiers

GB acc	Description
AA406467	zinc finger protein, X-linked
AA447835	small proline-rich protein 1B (cornifin)
T57959	zinc finger protein 268
AA043334	486544
H17047	zinc finger protein 133 (clone pHZ-13)
H62985	small inducible cytokine A4 (homologous to mous
AA425602	Human POM-ZP3 mRNA, complete cds
AA425102	small inducible cytokine A2 (monocyte chemotac
W16724	ESTs, Highly similar to MLL-AF4 der(11) fusion p
H29484	Sjogren syndrome antigen B (autoantigen La)
AA088564	zinc finger protein 38 (KOX 25)
AA411407	signal recognition particle 19kD

One row for each clone or probe assayed: that is one “spot” in 2 colour arrays” and one “probeset” in affymetrix chips.

Gene Annotations

GB acc	Acc	UGCluster	Name	Symbol	LLID	Chromosor	Cytoband	SumFunc	GO
AA406467	AA406467	Hs.2074	zinc finger	ZFX	7543	X	Xp21.3		molecular f
AA447835	AA447835	Hs.1076	small prolir	SPRR1B	6699	1	1q21-q22		molecular f
T57959	T57959	Hs.425991	ESTs, Highly similar to Z268_HUMAN ZINC FINGER PROTEIN 268 (ZINC FINC						
AA043334	AA043334	Hs.164915	small nucle	SNAPC3	6619	9	9p22.2		molecular f
H17047	H17047	Hs.78434	zinc finger	ZNF133	7692	20	20p11.23-20p11.22		molecular f
H62985	H62985	Hs.75703	chemokine	CCL4	6351	17	17q12		molecular f
AA425602	AA425602	Hs.296380	POM (POM	POMZP3	22932	7	7q11.23	This gene appears to h	
AA425102	AA425102	Hs.303649	chemokine	CCL2	6347	17	17q11.2-q2	This gene i	molecular f
W16724	W16724	Hs.199160	myeloid/lyn	MLL	4297	11	11q23		molecular f
H29484	H29484	Hs.83715	Sjogren syı	SSB	6741	2	2q31.1		molecular f
AA088564	AA088564	Hs.155470	zinc finger	ZNF3	7551	7	7q22.1	May mediate transcript	

Information retrieved from different data banks for each gene(spot/probe)

Where are the data?

- By default the data are hidden.
- You can manage to see some or all clicking the button in the upperleft corner with the legend

  “click to display the data”
- Warning! The button calls one macro in

  `C:\Program Files\ArrayTools\Excel\…`

  But if you are in Spain it has to be changed to

  `C:\Archivos de Programa\ArrayTools\Excel…`
- You can do it yourself rightclicking the button and changing this in the “Assign Macro” option

log ratios

GB acc	svcc77	svcc78	svcc86	svcc104	svcc85	svcc84
AA406467	0.177719161	0.635135	0.117507	-0.059445	-0.344543	-0.29523
AA447835	0.483082891	-0.938371	-1.925212	-1.088192	0.404442	-1.47468
T57959	-0.466033399	-0.667869	-0.661471	-0.949109	-1.118599	-0.951252
AA043334	-0.361635417	-0.52492	-0.180623	-0.975533	-1.337441	-1.108014
H17047	-0.180488467	0.467818	0.084874	-0.799491	-1.236151	-0.976088
H62985	0.617713928	0.186939	0.674088	1.229423	-0.529549	0.370547
AA425602	0.039528362	-0.568089	-0.431157	-0.65056	-1.235254	-0.905405
AA425102	-1.462129593	-1.363178	-0.753973	-0.962225	0.284074	-0.055937
W16724	-0.193975359	1.053527	-0.035172	-2.082636	-0.715087	-0.533485
H29484	-0.655079663	-1.870073	-0.550644	-0.837189	-1.01124	-1.155574
AA088564	-0.716671586	-0.725372	-0.720263	-0.899272	-1.529135	-1.312939

log2-transformed Red/Green ratios for two colour arrays or

Intensities for single- channel or Affymetrix chip data)

with the genes represented by the rows and the arrays by columns

2 & 3. Preprocessing steps: Filtering and Normalization

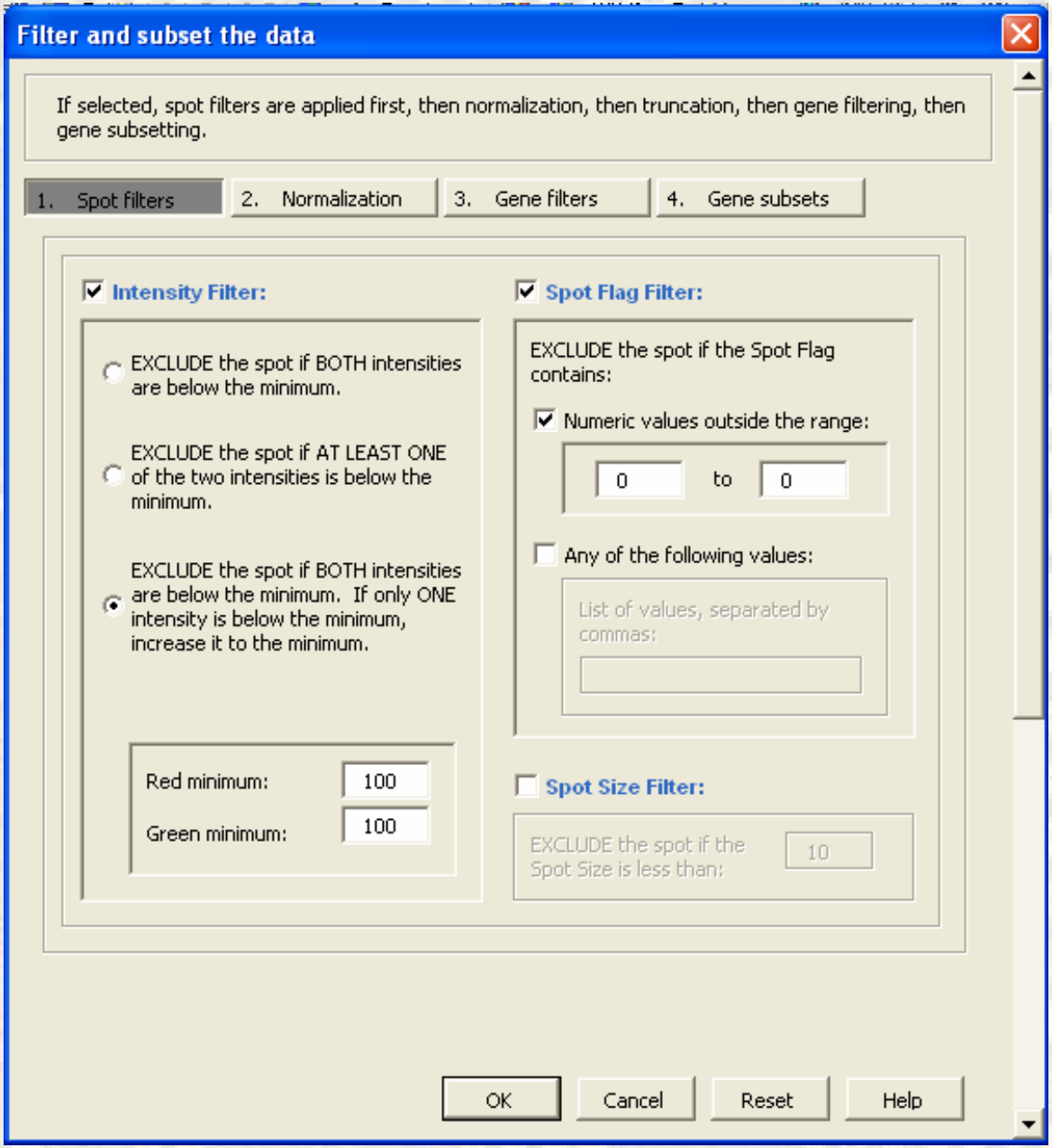
- After import/loading and before the analyis step data must be pre-processed.
- This may mean two type of actions:
  - *Filtering* is done to exclude bad spots or adjust intensities too low or too high to more reasonable values.
  - *Normalization* is done to correct for biases (systematic errors) due to technical reasons instead of biological variability.

2: Filtering spots & adjust signals

- We may filter the data on intensity by excluding values where both the red and green channels are less than 100.
- We may set the value of an intensity to the minimum in the event only one of the two channel intensities is below the minimum of 100.
- In addition, we may use the flag column imported with the data, and exclude intensities with a flag value not equal to 1.

Must we filter the data?

- Filtering is intended to remove spots whose images or signals were wrong due to different possible reasons
  - Small quantity of cDNA in the array
  - Errors during the scanning process
- Some people prefer not to filter to avoid eliminating good spots unintentionally.
- *In case of doubt be conservative and reduce the filter operation to the minimum.*



We filter following the tutorial's indications.

2998 genes pass the filtering criteria

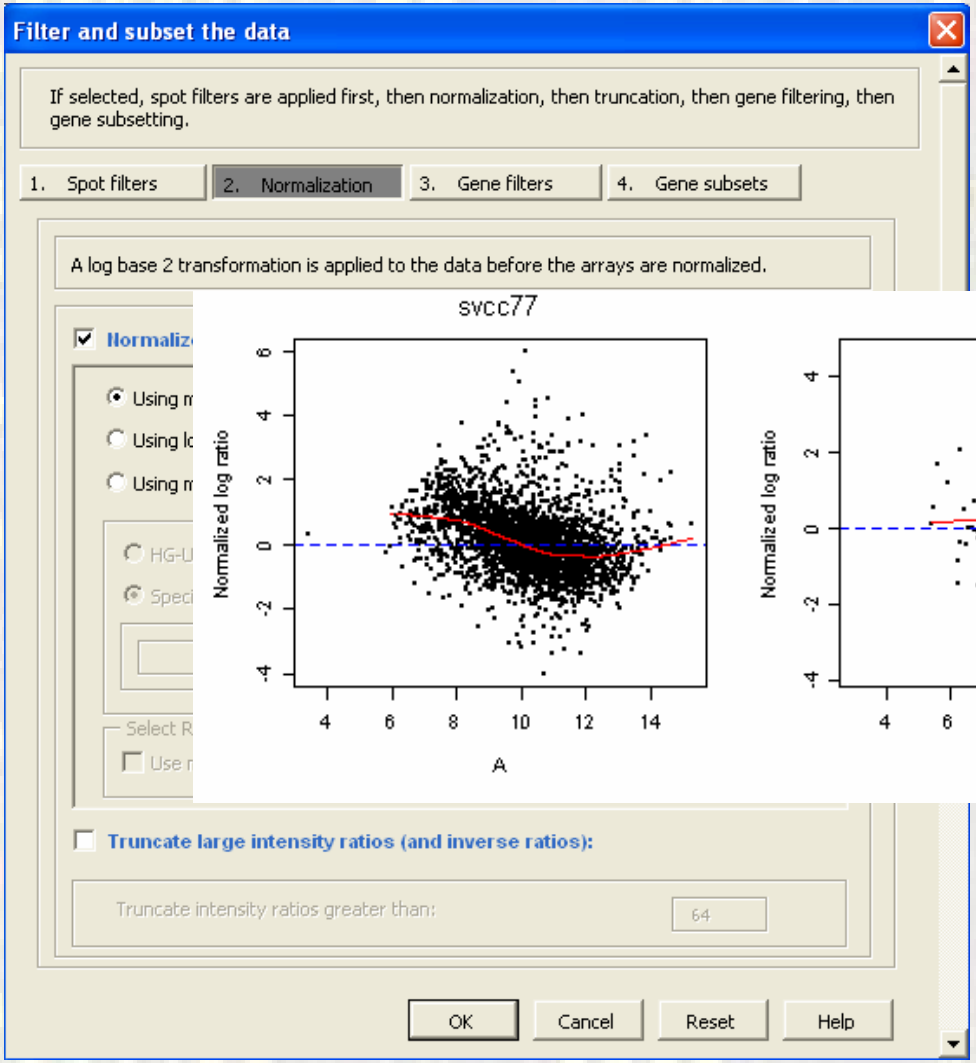
3. Normalization

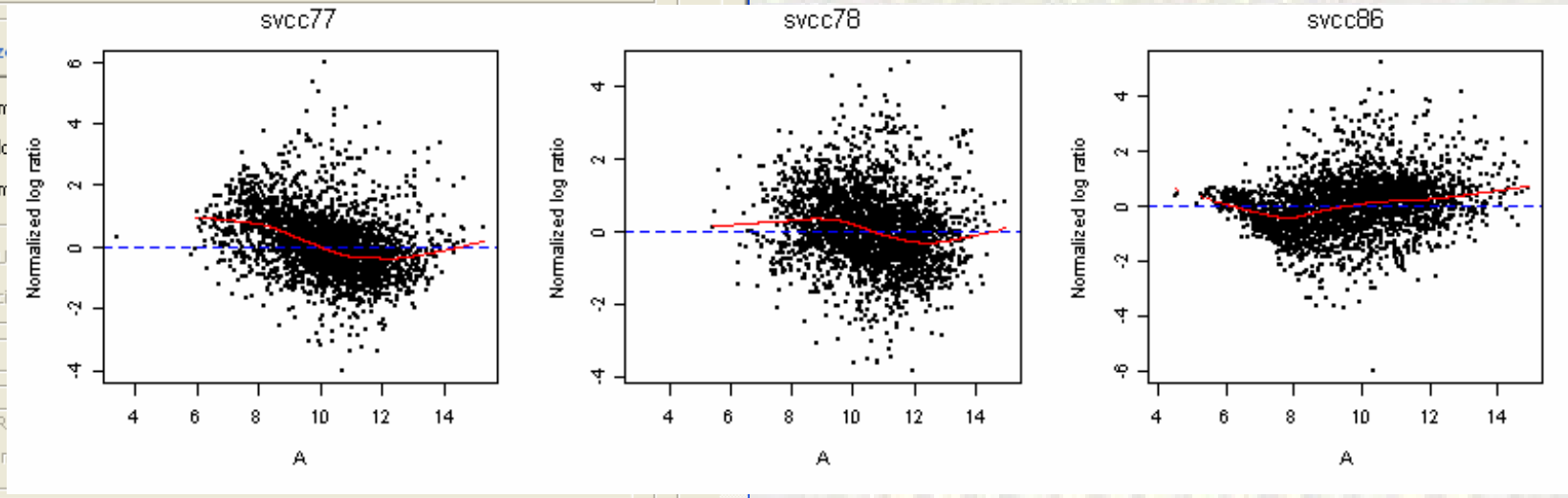
- A quick inspection of the data -e.g. MA plots- will show if normalization is needed
- First normalize the data subtracting the median log ratio of an array to all log ratios on that array.
- Later we will normalize the data by subtracting a non-linear transformation with the loess option.
  - No print-tip group information is available so it is not possible to perform print-tip normalization.
  - We will construct M-A plots to evaluate the results of each normalization option.

Is normalization necessary?

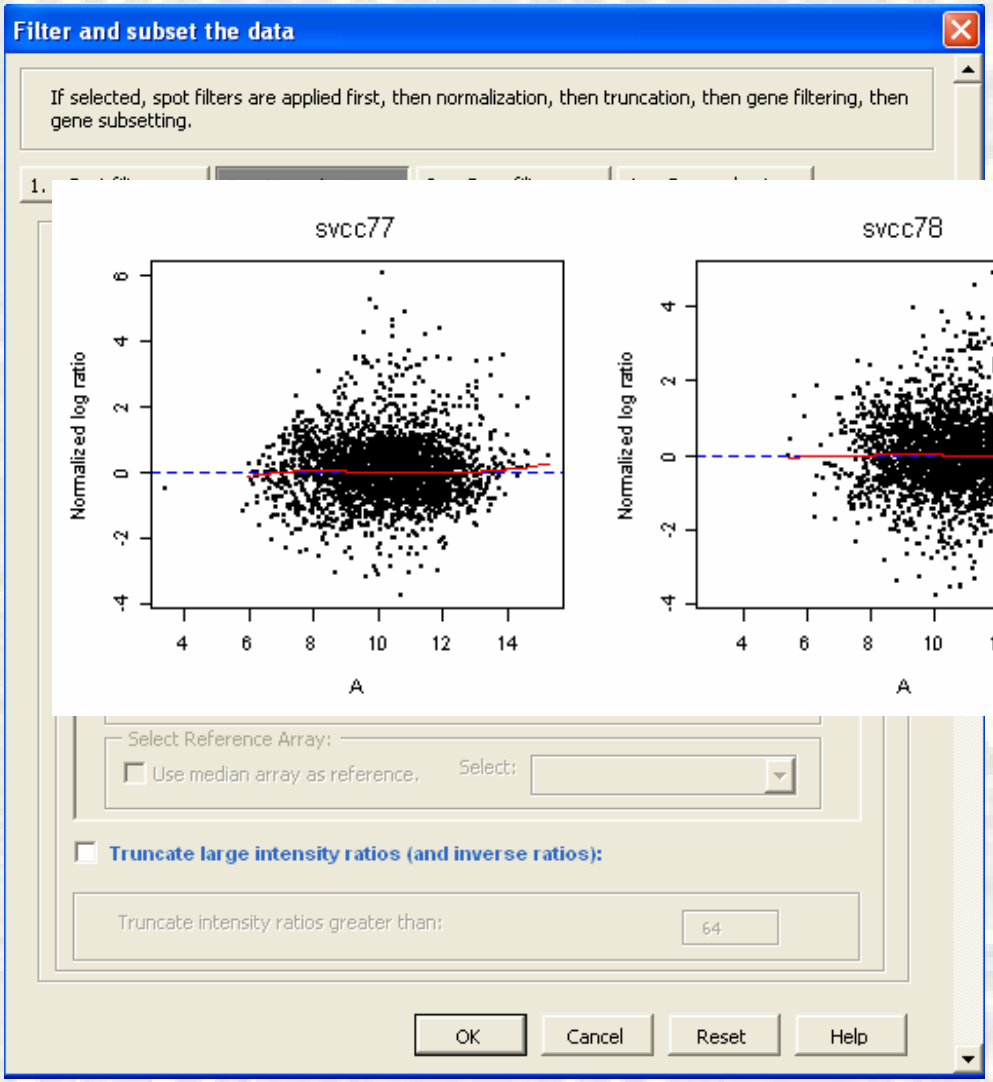
- MA plots can show if it is needed to normalize the data (it usually is)
- To draw an MA-plot go to:
  Array Tools → Plugins → M vs A plot
  - Asymetrical clouds, not centered around zero suggest the need for normalization.
  - Symetrical narrow clouds suggest that it can be omitted.

Median normalization

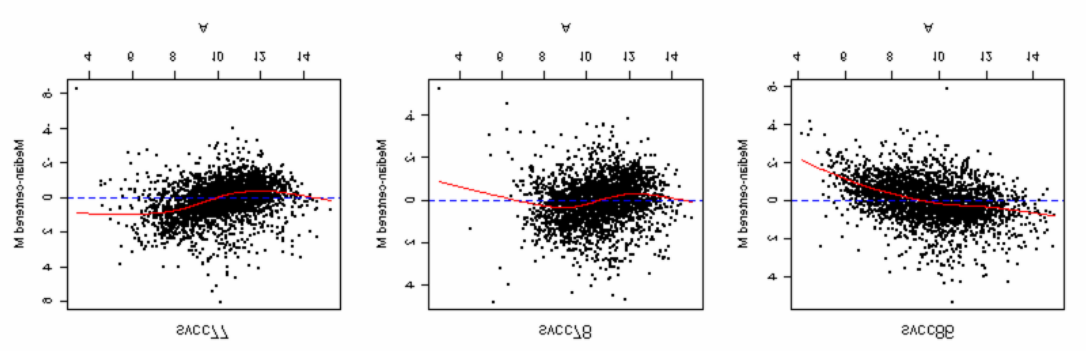




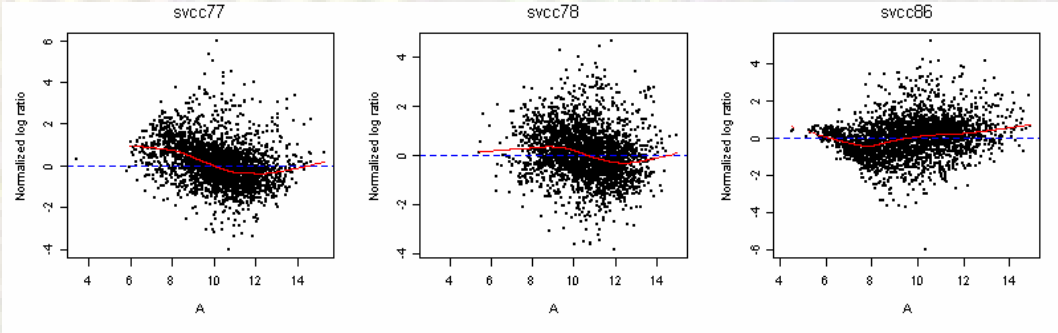
Loess normalization



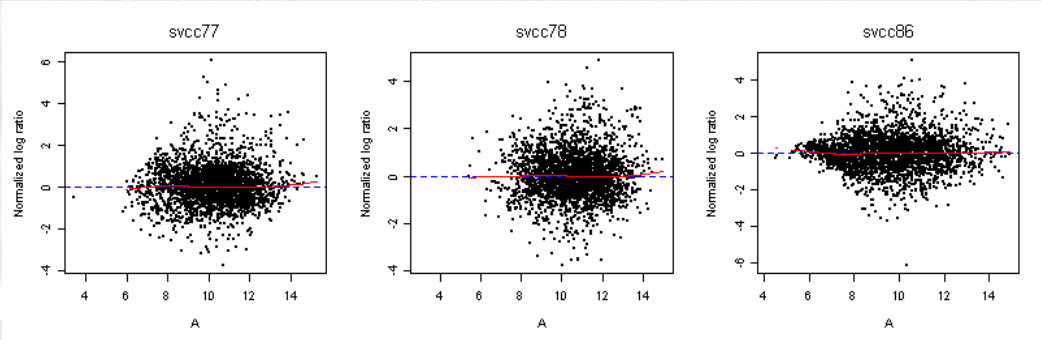
Before and after normalization



No normalization



Global median normalization

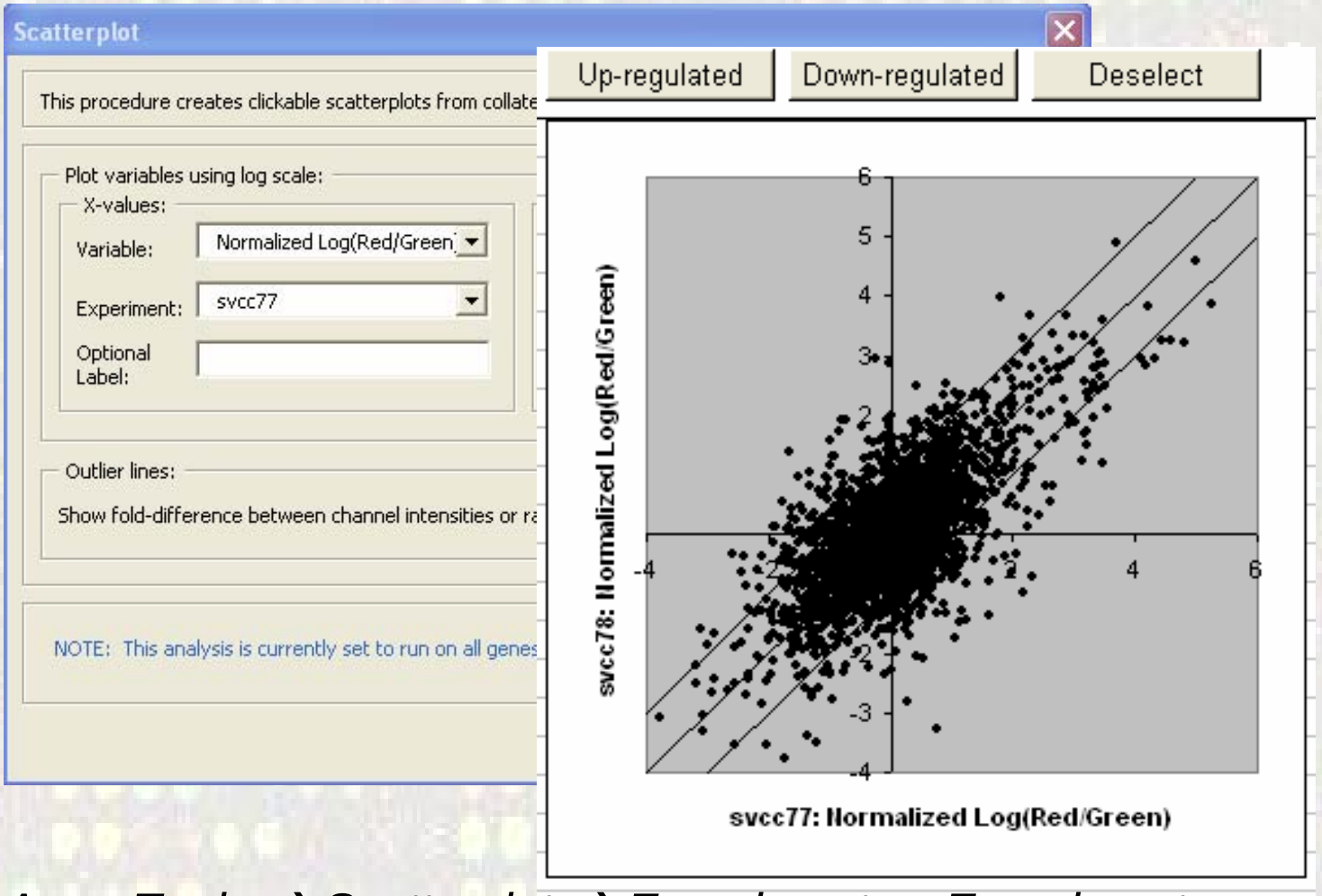


Global loess normalization

4. Finding differentially expressed genes

- Quick fold-change scatter plots can be used to make an inspection of up or down regulated genes in each experiment.
  - Useful to look at specific arrays.
  - Cannot be generalized.
- The best approach of course is to combine all samples and do a test of DE.

After vs Before scatterplot (svc77 vs svc78 slides)



Array Tools → Scatterplot → Experiment vs Experiment

Comparing visual checks

Gene Row	GB acc	Description
1440	AA001449	pleiotrophin (heparin binding growth factor 8, neurite gr
267	AA031513	matrix metalloproteinase 7 (matrilysin, uterine)
834	AA102670	Human GABA-A receptor pi subunit mRNA, complete
1881	AA423944	37 kDa leucine-rich repeat (LRR) protein
383	AA434024	770355
1566	AA434369	Human mRNA for KIAA0183 gene, partial cds
2060	AA455338	glycophorin B (includes Ss blood group)
671	AA459308	elastin (supravalvular aortic stenosis, Williams-Beuren
2203	AA490694	hevin
854	H05445	growth associated protein 43
2264	H08933	46054
950	H23978	general transcription factor IIB
322	H60423	solute carrier family 17 (sodium phosphate), member 1
594	H72937	2,4-dienoyl CoA reductase
2275	H75547	Homo sapiens clk1 mRNA, complete cds
1132	H91815	fibrinogen, B beta polypeptide
626	N26562	melan-A
862	N48137	glycophorin E
186	R10284	hyaluronan-mediated motility receptor (RHAMM)
92	R15708	insulin-like 4 (placenta)
757	R38201	opioid-binding protein/cell adhesion molecule-like
856	R89567	195340
2090	T73468	glutathione S-transferase A2
1714	T74819	85093

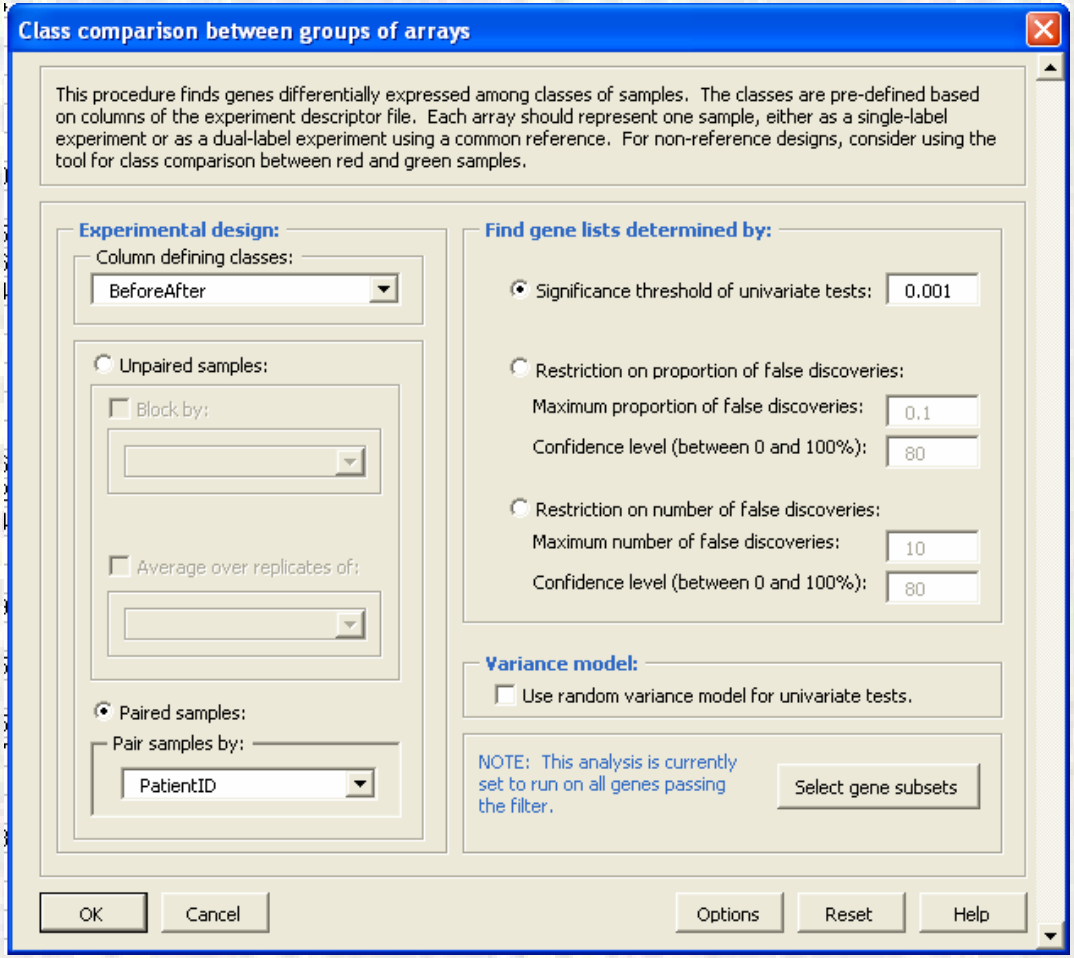
	Gene Row	GB acc	Description
1			
2	1132	H91815	fibrinogen, B beta polypeptide
3	398	H93328	Human putative cyclin G1 intera
4	394	R73003	solute carrier family 16 (monoc
5	1314	T73031	cytochrome P450, subfamily IIA
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			

- Te list of genes up-regulated before and after chemotherapy is not the same for patients 10 and 100

4.2 Class comparison tests

- A test for differential gene expression between pre and post chemotherapy can be done using a paired t-test.
- In order to avoid depending on normality assumptions p-values can be computed using a permutation approach.
- The number and proportion of false discoveries must be controlled. It can also be estimated

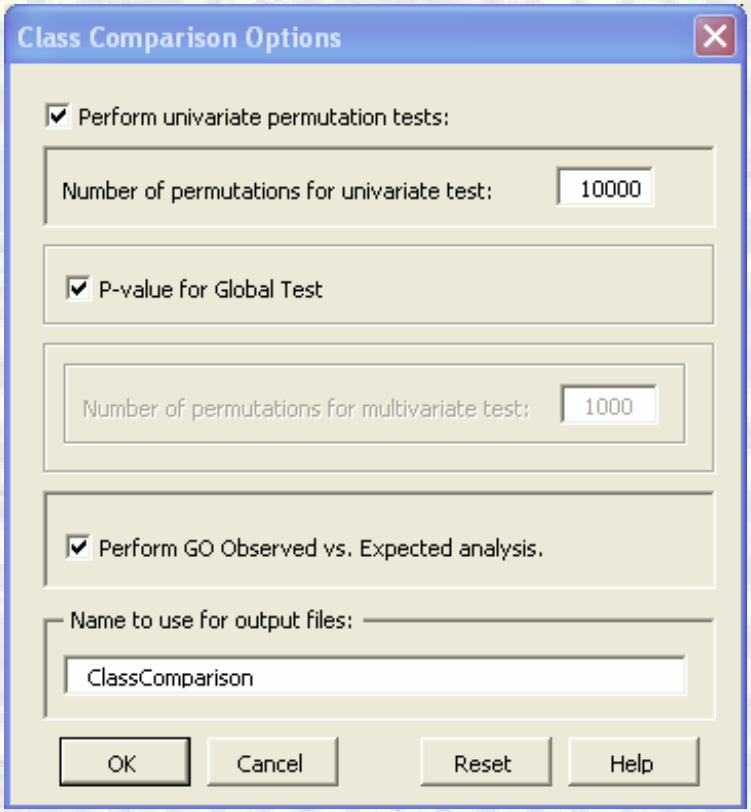
Class comparison: Select test



- There are several criteria to select genes
- But only one can be applied each time
- A threshold based on p-values is used in the example

Array Tools → Class Comparison → Between groups of arrays

Class comparison: Set options



- Using permutation test avoids having to do normality assumptions.
- Global test indicates the probability of selecting the genes finally chosen if there were no real differences.
- GO obs. vs exp.
  - can be used to find which functional classes appear to be enriched in the set of selected genes
  - Highlights functional relevant classes perhaps related to important biological processes acting on the experiment in this situation.

Array Tools → Class Comparison → Options

Results

- The analysis results are written to a file "ClassComparison.html"
- It contains
  - Description of the problem
  - Summary of Results
  - Genes which discriminate among classes
  - [*Optional*] 'Observed v. Expected' table of GO classes

Results (1): Summary

Description of the problem:

Number of classes: 2
Number of genes that passed filtering criteria: 2998
Type of univariate test used: Paired T-test
Column of the Experiment Descriptors sheet that defines class variable : BeforeAfter
Permutation p-values for significant genes were computed based on 10000 random permutations

Nominal significance level of each univariate test: 0.001

Summary of Results:

Number of genes significant at 0.001 level of the univariate test: 28
Global test: probability of getting at least 28 genes significant by chance (at the 0.001 level) if there are no real differences between the classes: 0

Results (2): List of genes

Genes which discriminate among classes:

Table 1 - Sorted by p-value of the univariate test.
The first 28 genes are significant at the nominal 0.001 level of the univariate test

	Parametric p-value	FDR	Permutation p-value	Geometric mean of ratios (class AF /class BE)	GB acc	Anno-tations	Description
1	2e-07	0.0005996	0	3.124	AA478553	Info	dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2)
2	1.1e-06	0.0016489	0	1.908	W96134	Info	Jun activation domain binding protein
3	5.3e-06	0.0042721	0	1.972	AA418077	Info	GTP-binding protein overexpressed in skeletal muscle
4	5.7e-06	0.0042721	0	1.743	AA455210	Info	platelet-derived growth factor receptor-like
5	1.53e-05	0.0085443	0	2.031	AA133129	Info	transcription elongation factor B (SIII), polypeptide 3 (110kD, elongin A)
6	1.71e-05	0.0085443	1e-04	2.038	T82477	Info	Duffy blood group
7	2.51e-05	0.0097935	0	1.79	AA287695	Info	701231
8	2.76e-05	0.0097935	0	1.987	H21041	Info	activating transcription factor 3
9	2.94e-05	0.0097935	0	2.032	AA598794	Info	connective tissue growth factor
10	4.53e-05	0.0135809	0	2.28	AA167222	Info	collagen, type XIV, alpha 1; undulin
11	7.66e-05	0.0207323	1e-04	0.679	H67349	Info	collagen, type IV, alpha 4
12	8.94e-05	0.0207323	2e-04	1.488	AA489234	Info	cytokine-inducible kinase
13	8.99e-05	0.0207323	0	1.431	H39192	Info	protein kinase mitogen-activated 7 (MAP kinase)
14	0.0001231	0.026361	0	0.825	AA598758	Info	tumor rejection antigen (gp96) 1
15	0.0002157	0.0431112	4e-04	1.375	AA464042	Info	collagen, type VI, alpha 2
16	0.0002327	0.0436022	1e-04	1.457	AA425139	Info	X-ray repair complementing defective repair in Chinese hamster cells 1
17	0.0002847	0.0502077	2e-04	1.508	AA486082	Info	serum/glucocorticoid regulated kinase
18	0.0003164	0.0526982	2e-04	1.326	R38383	Info	ESTs, Highly similar to TRISTETRAPROLINE [H.sapiens]
19	0.0003481	0.0549265	2e-04	0.61	T58146	Info	MHC class I region ORF
20	0.0004838	0.071024	7e-04	1.375	AA418670	Info	jun D proto-oncogene
21	0.0004975	0.071024	2e-04	0.57	N49629	Info	diubiquitin
22	0.000526	0.0716795	6e-04	1.187	R48232	Info	polycystic kidney disease (polycystin)-like
23	0.0007124	0.0921885	2e-04	1.434	R66310	Info	peptidylglycine alpha-amidating monooxygenase
24	0.000738	0.0921885	9e-04	1.961	T72581	Info	matrix metalloproteinase 9 (gelatinase B, 92kD gelatinase, 92kD type IV collagenase)
25	0.0008774	0.1005936	6e-04	0.829	H05563	Info	Human mRNA for KIAA0182 gene, partial cds
26	0.00088	0.1005936	5e-04	1.543	T62179	Info	FBJ murine osteosarcoma viral oncogene homolog B
27	0.0009385	0.1005936	8e-04	1.616	AA423944	Info	37 kDa leucine-rich repeat (LRR) protein
28	0.0009395	0.1005936	0.0012	1.927	AA486838	Info	secreted frizzled-related protein 4

Results (3): GO observed vs expected

'Observed v. Expected' table of GO classes and parent classes, in list of 28 genes shown above:

Only GO classes and parent classes with at least 5 observations in the selected subset and with an 'Observed vs. Expected' ratio of at least 2 are shown.

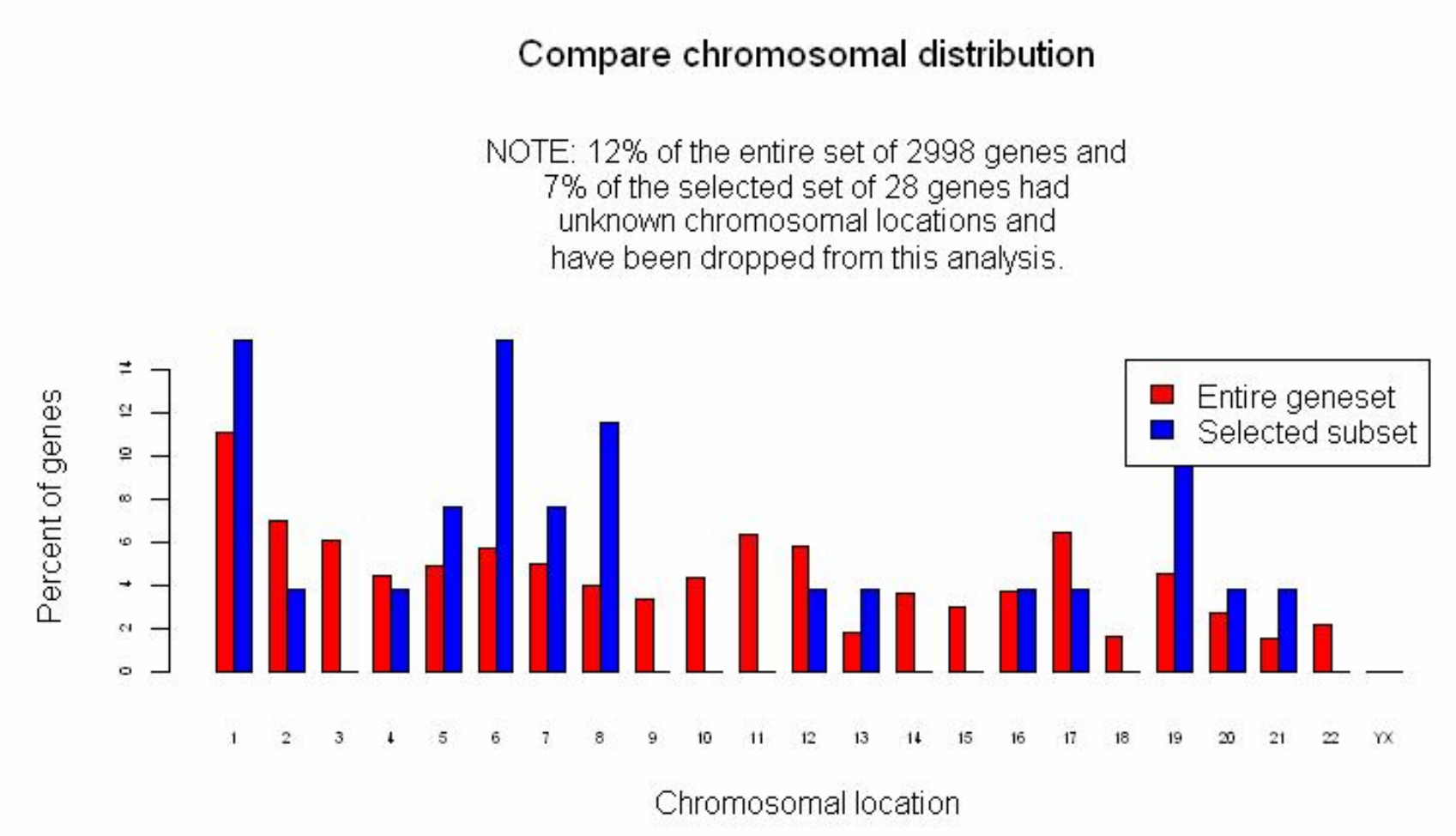
Cellular Component

GO id	GO classification	Observed in selected subset	Expected in selected subset	Observed/ Expected
0005578	extracellular matrix (sensu Metazoa)	5	0.82	6.13
0031012	extracellular matrix	5	0.83	6.06
0005576	extracellular region	5	2.33	2.14

Biological Process

GO id	GO classification	Observed in selected subset	Expected in selected subset	Observed/ Expected
0006811	ion transport	5	1.23	4.07
0007155	cell adhesion	5	1.47	3.41
0009607	response to biotic stimulus	6	2.46	2.43
0006950	response to stress	6	2.53	2.37
0050896	response to stimulus	10	4.6	2.18

Results (4): Chromosomal distribution



Results (4): Some extra info.

Date and time of the analysis:

Name of the project file: Perou.xls
Time of the analysis: Sat Mar 17 18:55:32 2007
BRB-ArrayTools Version: 3.5.0 - Patch_1 (March 2007)

Filtering parameters:

- **Spot Filters:**

If RED intensity is below 100, increase it to 100. If GREEN intensity is below 100, increase it to 100. Exclude the spot if RED intensity is below 100 AND GREEN intensity is below 100 .
Exclude the spot if the Spot Flag contains numeric values outside the range: 0 to 0

- **Normalization:**

Normalize (center) each array using lowess smoother.

- **Gene Filters: OFF**

- **Gene Subsets: OFF**

To learn more …

- Analysis of Gene Expression Data Using BRB-Array Tools
  Richard Simon, Amy Lam, Ming-Chung Li, Michael Ngan, Supriya Menenzes, Yingdong Zhao
  Cancer Informatics 2:11-17, 2007.
- A Tutorial on Data Analysis Using BRB-ArrayTools version 3.5
  Supriya Menezes
  NIH CIT course on BRB-ArrayTools, October 2006.