

Tools and Algorithms in Bioinformatics

GCBA815, Fall 2013

Week7: Microarray data analysis

BRB tools

Babu Guda & Xiaosheng Wang

Department of Genetics, Cell Biology and Anatomy

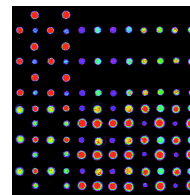
University of Nebraska Medical Center

10/11/2013

GCBA 815

Microarrays

- GeneChips
 - Contain short (25-mer) oligonucleotides as probes
 - About 400,000 different DNA spots per chip
 - Space for a lot of controls and multiple spots (16-20) per gene
 - Mismatched oligos for background correction
- Spotted Arrays
 - Concentrated DNA is prepared and spotted on glass slides
 - Much cheaper than GeneChips
 - Designed to contain only the genes of interest
 - Usually limited to about 15,000 spots per slide
 - Limits controls and duplicates for each gene

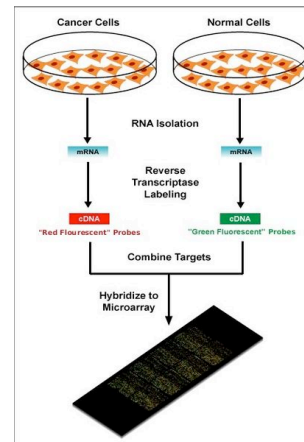


10/11/2013

GCBA 815

Two-channel vs one-channel detection

- Two-channel (or two-color)
 - Typically used to compare expression in two different treatments
 - Two types of cDNA labeling dyes are used
 - Cy3- has emission at 570nm (corresponds to green)
 - Cy5- has emission at 670nm (corresponds to red)
 - The two Cy-labeled cDNA samples are mixed and hybridized to the a microarray
 - Relative intensities are used detect up or down-regulated genes



10/11/2013

GCBA 815

Two-channel vs one-channel detection

- Single-channel (one-color)
 - Uses a single dye to measure intensity for each probe
 - Multiple treatments are hybridized to multiple arrays and the intensities are compared to determine up or down-regulated genes
 - One bad quality sample does not affect the other array results (which is a problem in two-channel arrays)
 - Results are easily comparable to arrays from different experiments
 - Requires twice as many microarrays compared to two-channel arrays

10/11/2013

GCBA 815

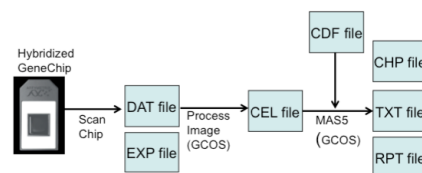
Experimental design

- Controls (RNA spike-ins)
- Technical replicates
 - Multiple aliquots of the same sample are run separately to account for technical variation
- Biological replicates
 - Multiple independent samples of the same kind are run separately to account for sample to sample variation
- Multiple spots for each gene on a single array to measure statistical significance.

10/11/2013

GCBA 815

Affymetrix File Types



- DAT file
 - Contains raw (TIFF) optical image of the hybridized chip
- EXP file
 - Text file with experimental details
- CEL file
 - Processed DAT file (with intensity/position values)
- CDF file
 - Described layout of chip (provided by Affymetrix)
- CHP file
 - Results created from CEL and CDF file
- TXT file
 - CHP file in text format
- RPT file
 - Report file with QC information

10/11/2013

GCBA 815

Basic statistics

- Raw intensities
 - Pre-processing, background correction
 - Transformed into expression values (called normalization)
 - Use algorithms such as MAS5, RMA/GCRMA, etc.
- Fold change
 - Ratio of normalized intensities of experimental/control
- Log(intensity) or log(ratio)
 - Brings the differential expression values from multiplicative scale to additive scale
- T-test (Gosset 'student' t-test)
- P-value (obtained from a test statistic such as t-test)

10/11/2013


GCBA 815

Log transformations

Normalized Intensity	Natural Log (ln)	Log ₂	Log ₁₀
1	0	0	0
10	2.30	3.32	1
20	3	4.32	1.3
100	4.61	6.62	2
1000	6.91	9.97	3
10000	9.21	13.29	4
50000	10.82	15.61	4.7
100000	11.51	16.61	5

10/11/2013

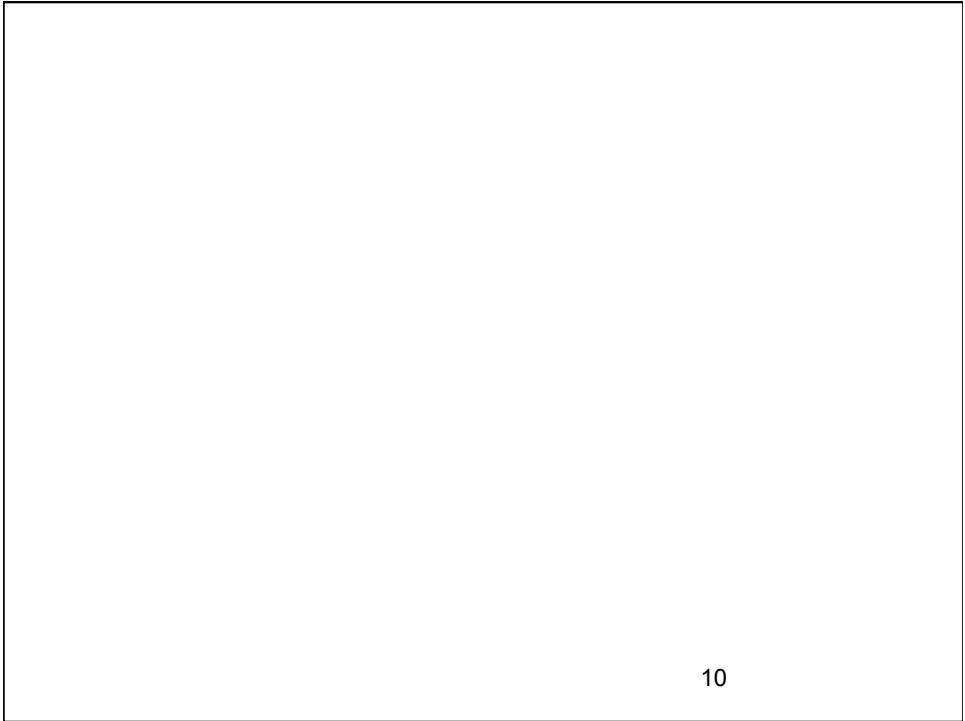
GCBA 815



Public Repositories of Microarray Datasets

- NCBI's GEO (Gene Expression Omnibus)
 - GEO data sets
 - Original expression datasets
 - Curated gene expression datasets
 - Cluster tools and differential expression queries
 - Geo Profiles
 - Contains gene expression profiles for each gene
- EBI's Array Express

10/11/2013 GCBA 815



10

Microarray Data Analysis Using BRB-ArrayTools

Xiaosheng Wang

11

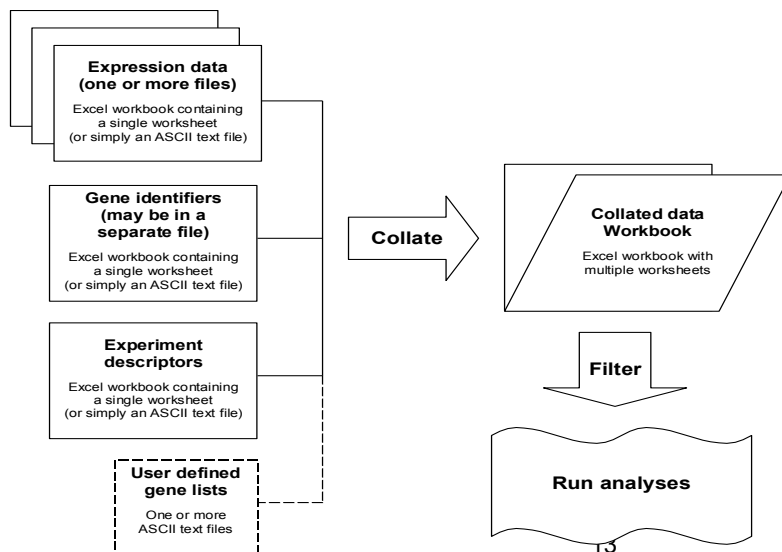
BRB-ArrayTools

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.

12

Data input to BRB-ArrayTools



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

		Array data block #1			Array data block #2			Array data block #3			
A	B	C	D	E	F	G	H	I	J	K	
Wellid	Clone	Description	Red_1	Green_1	Flag_1	Red_2	Green_2	Flag_2	Red_3	Green_3	Flag_3
2	600001	IMAGE:604856	adhesion selectin B Mrr	21363	13268	0	19674	11840	0	11938	4670
3	600002	IMAGE:619876	adhesion VCAM-1 Mm.1	16895	11908	0	45073	30279	0	16194	7691
4	600003	IMAGE:442991	adhesion ELAM Mm.21	3823	2511	0	8238	3657	0	6674	1962
5	600004	IMAGE:615729	adhesion integrin5b	11277	5950	0	11045	6706	0	7000	3879
6	600005	IMAGE:522319	adhesion integrin a5 Mm	8979	3402	0	12431	3497	0	7660	1871
7	600006	IMAGE:576194	adhesion integrin B1	17472	12238	0	14281	10961	0	14337	6918
8	600007	IMAGE:533853	adhesion thrombospondin	14204	6937	0	14476	4305	0	9043	2321
9	600008	IMAGE:476623	adhesion ICAM Mm.394	17872	9622	0	22588	12239	0	11048	5572
10	600009	IMAGE:539626	adhesion integrin a4 Mm	30205	15216	0	43000	14654	0	19379	5698
11	600010	IMAGE:678744	adhesion integrin a2	18122	9274	0	21378	10640	0	12177	4697
12	600011	IMAGE:679592	adhesion integrin C8 Mr	49522	25469	0	53653	21495	0	30237	8461
13	600012	IMAGE:426454	adhesion integrin E7 Mr	38276	17583	0	40191	15761	0	21315	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	1305
17	600016	IMAGE:639320	adhesion selectin P	3813	3368	0	4176	3991	0	3841	2332
18	600017	IMAGE:677203	adhesion selectin E Mrr	5201	3209	0	5314	2059	0	2305	7709
19	600018	IMAGE:672927	adhesion SOM1	8793	4038	0	13467	4956	0	7651	1788
20	600019	IMAGE:535792	adhesion cadherin 5 Mm	9162	15130	0	7701	12335	0	3214	5331
21	600020	IMAGE:473150	adhesion thrombospondin	16010	5794	0	20450	7963	0	10764	3165
22	600021	IMAGE:639878	adhesion integrin a9	3649	3065	0	4291	3198	0	1911	1363
23	600022	IMAGE:521894	adhesion fibronectin	3115	2737	0	7156	7223	0	6637	1868
24	600023	MP-1B1	adhesion integrin B1	3139	1770	0	2900	822	0	1417	505

10/11/2013 15 GCBA 815 15

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.

16

Gene identifiers

Two examples of a gene identifier file

The top screenshot shows an Excel spreadsheet titled 'Geneids.xls' with the following data:

	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

The bottom screenshot shows an Excel spreadsheet titled 'Gene_identifiers.xls' with the following data:

	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=Tyrosin	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	DSIP1	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	BRE2	2p22.3-2p21	

17

Experiment (Array) descriptors

- An experiment descriptors file describes the samples used for each array, and is mandatory.
- After the header row, each row in this file represents one array or sample, and each column represents one descriptor variable.
- First column contains array id, subsequent columns contain descriptions, phenotype class labels, patient outcome, and other sample or experiment information.
- 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.

18

Experiment descriptors

Describes the samples used for each array

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

19

Automatic data importers

- General format data: The data import wizard can be used to guide you through the specification of the data components.
- Affymetrix data: Automatically imports data by searching for “Probe Set Name”, “Signal” (or “Avg Diff”), and “Detection” (or “Abs_Call”) column header labels.
- For importing Affymetrix CEL files, go to the following menu items: (Data Import Wizard), find a data folder containing the .CEL files, and provide an Experiment Descriptors file. Gene identifiers will be imported automatically from the BRB server.
- 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.

20

Data filtering options

- Single-Channel

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.

21

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.

22

Data filtering options

Normalization and truncation

- 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)

23

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

24

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.

25

Data filtering options

Gene subsets

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

26

Some important analysis tools

- **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.

27

Finding Genes

- Comparing classes (**Class Comparison**)

- Correlated with a quantitative trait (**Quantitative Trait Analysis**)

- Correlated with survival (**Survival Analysis**)

- Time Course Analysis (**Plug-in**)₂₈

Tools for Class Comparison

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

29

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
- Use the Class Prediction tool when the primary interest is to form a classifier to predict the class of new samples.

30

Components of Class Prediction

- C1. Feature(gene) selection
 - which genes will be included in the model.
- C2. Select model type.
 - choose prediction method (SVM, k-NN etc)
 - Fit the parameters for the model.
- C3. Evaluating the Classifier
 - Cross-validation

31

Cross-validating the classifier

- Leave-One-Out cross validation.
- K-Fold cross validation.
- +0.632 bootstrap cross-validation.

The screenshot shows a software interface for configuring cross-validation. It is divided into two main sections. The left section, titled "Cross-validation method:", contains three radio button options: "Leave-one-out validation" (which is selected), "10 - fold validation" (with a sub-section for "Repeated 1 times"), and "0.632 bootstrap validation". The right section contains a checkbox labeled "Do statistical significance test of cross-validated mis-classification rate." which is currently unchecked. Below this checkbox is a text label "Number of permutations for significance test of cross-validated mis-classification rate:" followed by a text input field containing the number "100".

32

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

33

Other tools

- Plugins: allows users to create their own tools by writing their own scripts written in the R language which can call C and Fortran programming.
- Methylation tools: find frequently methylated probes and correlate methylation with gene expression.
- Utility tools: annotate data, create gene lists, DrugBank information for significant genes.
- CGHTools: analyze copy number variation.

34