Analysis of microarray data

Claudio Altafini

SISSA

http://people.sissa.it/∼altafini

Analysis of microarray data

Introduction
- Analysis of microarray data
- Plots & statistics

Low level analysis of microarrays

Differentially expressed genes

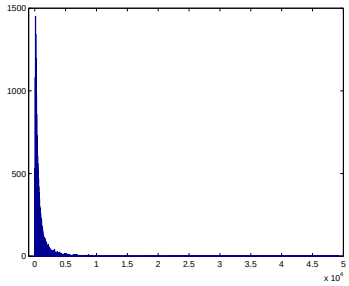
what do you do with microarray experiments?

1. low level analysis
   - from scanned images to gene expression
2. high level analysis (model-free)
   - identify differentially expressed genes
   - clustering
   - principal component analysis
   - ontological enrichment
3. modeling
   - reverse engineering
   - merging with a-priori information

Plots & statistics

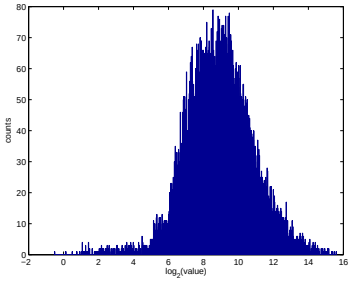
Introduction
Analysis of microarray data
Plots & statistics
Low level analysis of microarrays
Differentially expressed genes

- distribution of values: histogram



  - you do not see anything.....
  - many low values, a few very high values

- take the $\log_2$:



  - distribution is *almost* normal $\Longrightarrow$"log is good"
  - normal distribution: many statistical tests

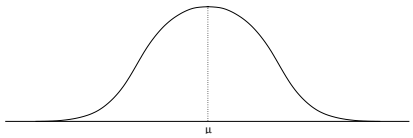
Plots & statistics

Normal distribution $\mathcal{N}(\mu,\ \sigma^2)$

Introduction
Analysis of microarray data
Plots & statistics
Low level analysis of microarrays
Differentially expressed genes

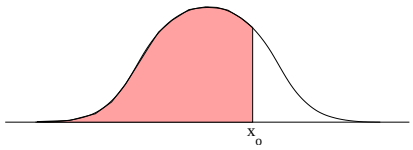
- probability density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- (cumulative) probability distribution

$$P(x \leqslant x_o) = \text{ area}$$

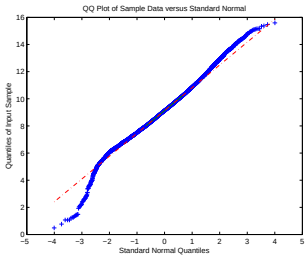


- standardization:

$$x \in \mathcal{N}(\mu,\ \sigma^2) \longrightarrow z = \frac{x-\mu}{\sigma} \in \mathcal{N}(0,\ 1)$$

  - z = z-score = standard normal distribution

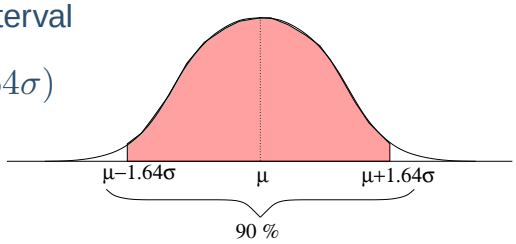
Plots & statistics



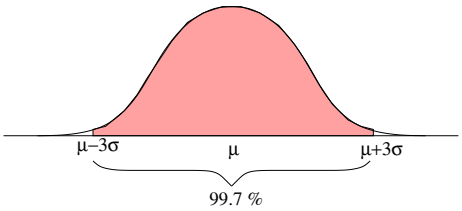
- to check normality:
  quantile-quantile plot
  (histogram vs normal distribution)

- confidence interval = how much we trust that a value lies in a given interval

$$P(\mu - 1.64\sigma < x < \mu + 1.64\sigma) = \text{area} = 0.90$$



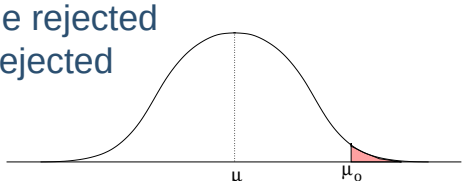
$$P(\mu - 3\sigma < x < \mu + 3\sigma) = \text{area} = 0.997$$



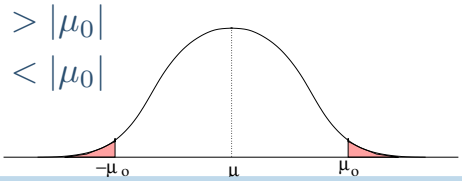
Plots & statistics

P-value

- given a sample distribution $X_1, \ldots, X_n$ of mean μ
- given a value μ_0
- Hypoteses: $\begin{cases} \text{null hypotesis} & H_0: \quad \mu > \mu_0 \\ \text{alternative hypotesis} & H_1: \quad \mu < \mu_0 \end{cases}$
- P-value = signifi cance value of the null hypotesis H_0, according to a test statistics
  - P-value high $\Longrightarrow H_0$ cannot be rejected
  - P-value low $\Longrightarrow H_0$ must be rejected
- typical choice of P-value: 0.05
- for $n > 30$ can assume
  $x_1, \ldots, x_n \in \mathcal{N}(\mu, \sigma) \implies$ P-value = area $P(x > \mu_0)$



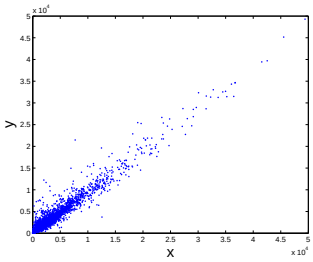
- two-sided P-value $\begin{cases} H_0: & |\mu| > |\mu_0| \\ H_1: & |\mu| < |\mu_0| \end{cases}$



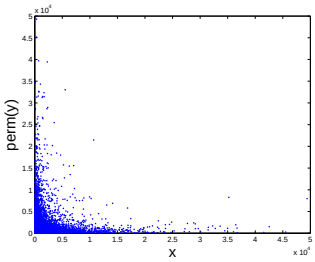
Plots & statistics

given 2 arrays $\mathbf{x}$, $\mathbf{y}$

- scatter plots

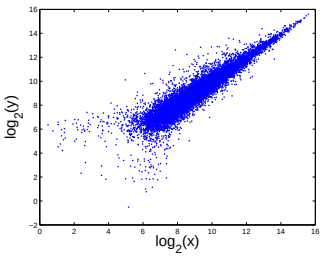


correlated

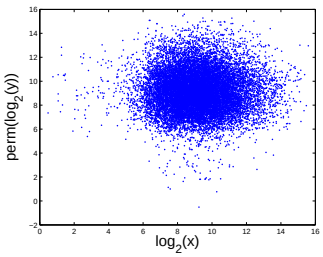


uncorrelated

- take log



correlated



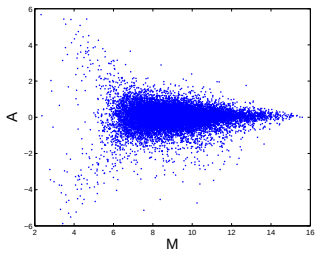
uncorrelated

Plots & statistics

- M-A plot: intensity vs ratio

$$A = \frac{\log_2(\mathbf{x} \cdot \mathbf{y})}{2} = \frac{\log_2(\mathbf{x}) + \log_2(\mathbf{y})}{2}$$

$$M = \log_2\left(\frac{\mathbf{x}}{\mathbf{y}}\right) = \log_2(\mathbf{x}) - \log_2(\mathbf{y})$$

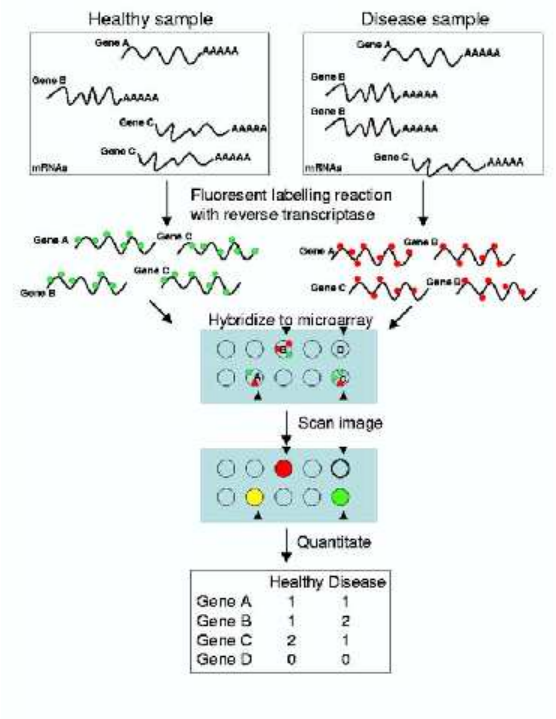


(Statistical) analysis of microarray data

Introduction

Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

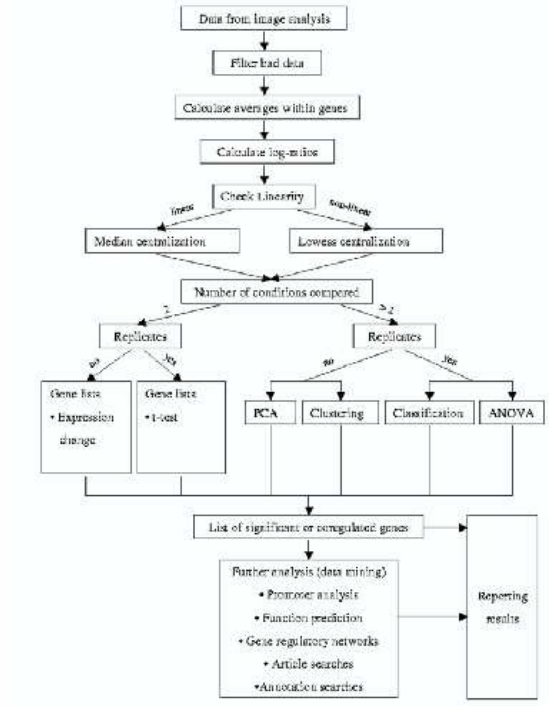


(Statistical) analysis of microarray data

Introduction

Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes



Microarray technologies

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

- 2 types of microarrays: { cDNA, Affymetrix GeneChips }
- cDNA
  - full length DNA clones
  - dual channel: (competitively) hybridized and labeled with different fluoresent dyes
    - Cy5: red-fluorescent dye
    - Cy3: green-fluorescent dye
  - measure only ratios of fluorescence intensities
  - cheap, and often custom made
  - low fidelity

Microarray technologies

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

- Affymetrix GeneChips
  - single channel: a single RNA is hybridized on the array
  - each gene is represented by a set of probes (11 - 20)
  - probe = 10 - 25 oligonucleotide pairs
  - probe pair
    - PM = Perfect Match of the desired sequence
    - MM = MisMatch, has a single nucleotide mismatch in the middle of the sequence
  - expensive ($\simeq$ 1000 € per slide)
  - more reproducible

Low level analysis

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

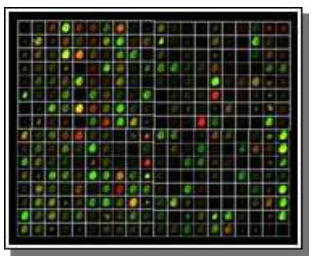
- 3 steps
  1. image analysis and background adjustement
     - read and interpret the image from the scanner
     - correct the signals for the background intensity
  2. normalization
     - manipulate the data to make measurement from different arrays compatible
  3. summarization
     - combine together multiple probe intensity for the same probeset to produce an expression value (only Affymetrix)
- References:
  - book on line “DNA Microarray Data Analysis” `http://www.csc.fi/molbio/arraybook/`
  - Affymetrix GeneChip Manual
  - Bioconductor project: `http://www.bioconductor.org`

Image analysis and background adjustement

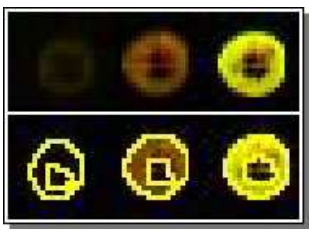
Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

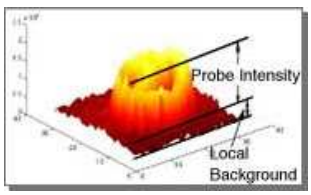
Differentially expressed genes

- on the scanned image place a grid around each spot



- extract a reading from the pixels that correspond to the spot





Example: cDNA

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

scanner reading details for a channel of a typical cDNA:

Column Title	Description
Block	the block number of the feature.
ID	the unique identifi er of the feature derived from the Array List
X	the X-coordinate in μm of the center of the feature-indicator associated with the feature
Y	the Y-coordinate in μm of the center of the feature-indicator associated with the feature
Dia.	the diameter in ţm of the feature-indicator.
F635 Median	median feature pixel intensity at wavelength #1 (635 nm).
F635 Mean	mean feature pixel intensity at wavelength #1 (635 nm).
F635 SD	the standard deviation of the feature pixel intensity at wavelength #1 (635 nm).
B635 Median	the median feature background intensity at wavelength #1 (635 nm).
B635 Mean	the mean feature background intensity at wavelength #1 (635 nm).
B635 SD	the standard deviation of the feature background intensity at wavelength #1 (635 nm).
% > B635 + 1 SD	the percentage of feature pixels with intensities more than one standard deviation above the background pixel inter
% > B635 + 2 SD	the percentage of feature pixels with intensities more than two standard deviations above the background pixel inte
F635 % Sat.	the percentage of feature pixels at wavelength #1 that are saturated.
F1 Median - B1	the median feature pixel intensity at wavelength #1 with the median background subtracted.
F1 Mean - B1	the mean feature pixel intensity at wavelength #1 with the median background subtracted.
SNR 1	the signal-to-noise ratio at wavelength #1, defi ned by (Mean Foregr. 1- Mean Backgr. 1) / (St. dev. of Backgr. 1)
F1 Total Intensity	the sum of feature pixel intensities at wavelength #1

Example: Affymetrix

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

- for the image reading part: you normally “trust” the standard image reading algorithms of Affymetrix GeneChip software
- files
  - `*.dat`: data file, scanned image of the probe array
  - `*.cel`: cell intensity file, contains a single intensity values for each probe
- each probe: 64 pixels
- disregard border pixels
- take 75% as cell’s value

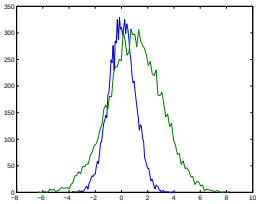
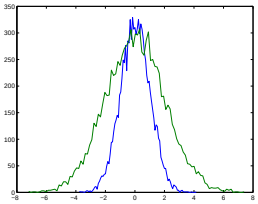
Normalization

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

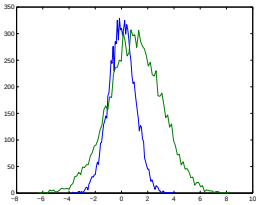
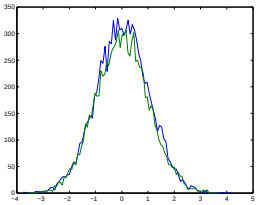
- global scaling (i.e. shift of the mean)

$$\mathbf{y} = \mathbf{y} - \bar{\mathbf{y}} + \bar{\mathbf{x}}$$

- standardization (i.e. shift the mean + same standard dev.)

$$\mathbf{y} = (\mathbf{y} - \bar{\mathbf{y}} + \bar{\mathbf{x}}) \frac{s_{\mathbf{x}}}{s_{\mathbf{y}}}$$

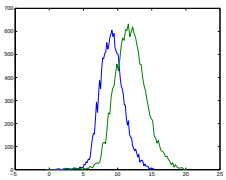
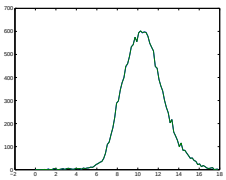



Normalization

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

- quantile normalization

1. sort each array
2. take average on the ranking
3. reassign the averages to the corresponding ranks

$\Longrightarrow$probability distributions are exactly the same

exp 1	exp 2
50	1000
400	600
200	100

$\Longrightarrow$

exp 1	exp 2
50	100
200	600
400	1000

$\Longrightarrow$

average
75
400
700

$\Longrightarrow$

exp 1	exp 2
75	700
700	400
400	75

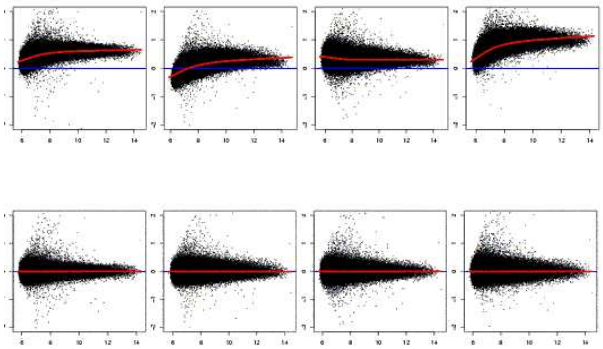
Normalization

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

- lowess normalization
  1. fit a smooth curve to the MA plot
  2. “straigten up” the curve
  3. retrasnform the M, A, data into $\mathbf{x}$, $\mathbf{y}$

  $\Longrightarrow$correct only when there is an intensity dependent bias
  $\Longrightarrow$nonlinear normalization



Affymetrix preprocessing: MAS 5.0

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

- MAS = MicroArray Suite is the “official” Affymetrix processing software
- BACKGROUND ADJUSTEMENT
  - chip is divided into a grid of rectangular regions (default: 16)
  - on each region: background is computed using the lowest 2 % of probe intensities
  - each probe intensity (both PM and MM) is adjusted based on a weighted average of these background values
- NORMALIZATION:
  - scaling
    1. global scaling (when there are few changes in gene expression among the arrays)
    2. scaling according to “housekeeping genes” (when there are many changes among the arrays)
  - robust normalization: further compensation

Affymetrix preprocessing: MAS 5.0

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

- SUMMARIZATION:
  - for each PM/MM probe pair compute the discrimination score:
$$R = \frac{PM - MM}{PM + MM}$$
  - if $R > 0.015 \Longrightarrow$probe pair is voting for the *presence* of the transcript
  - if $R < 0.015 \Longrightarrow$probe pair is voting for the *absence* of the transcript
  - *detection p-value*: perform a One-sided Wilcoxon's signed rank test on all probe pairs
  - p-value is used to assign a detection flag to a transcript:

p-value	flag	detection
p-value < 0.04	**P**	transcript is PRESENT
0.04 < p-value < 0.06	**M**	trascpript is MARGINAL
p-value > 0.06	**A**	trascpript is ABSENT

Affymetrix preprocessing: MAS 5.0

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

- SIGNAL QUANTIFICATION
  - assigns the relative level of expression to the transcript
  - computed as a weighted mean using the One-Step Tukey's Biweight Estimate
  - all pairs contribute to the estimate, with various corrections (e.g. MM>PM is physiological nonsense; probe pairs closer to median have heavier weight; etc.)

- scaling can precede or follow summarization

RMA preprocessing

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

- RMA = Robust Multichip Average
- more recent preprocessing method for Affymetrix Chips
- BACKGROUND ADJUSTEMENT
  - uses only PM (reason: MM values are strongly dependent on PM)
  - PM probes are modeled as sums of Gaussian noise $\mathcal{N}(\mu, \sigma^2)$ + exponential signal component $\mathrm{Exp}(\alpha)$
- NORMALIZATION: quantile
- SUMMARIZATION: median polish
  - aims at centralizing both columns (chips) and rows (genes) medians to 1
    1. compute row median and subtract it, yielding row median =1
    2. do the same with columns
    3. repeat until it converge

cDNA preprocessing

Introduction
Low level analysis of microarrays
- Microarray technologies
- Low level analysis
- Image analysis
- Normalization
- MAS 5.0
- RMA
- cDNA preprocessing

Differentially expressed genes

- similar algorithms
- dual channel: only comptetitive hybridization $\Longrightarrow$ratios
- single spot per gene: less accurate measure

Ratio of Medians	the ratio of the median intensities of each feature for each wavelength, with the median background subtracted.
Ratio of Means	the ratio of the arithmetic mean intensities of each feature for each wavelength, with the median background subtra
Median of Ratios	the median of pixel-by-pixel ratios of pixel intensities, with the median background subtracted.
Mean of Ratios	the geometric mean of the pixel-by-pixel ratios of pixel intensities, with the median background subtracted.
Ratios SD	the geometric standard deviation of the pixel intensity ratios.
Rgn Ratio	the regression ratio of every pixel in a 2-feature-diameter circle around the center of the feature.
Rgn R^2	the coeffi cient of determination for the current regression value.
F Pixels	the total number of feature pixels.
B Pixels	the total number of background pixels.
Sum of Medians	the sum of the median intensities for each wavelength, with the median background subtracted.
Sum of Means	the sum of the arithmetic mean intensities for each wavelength, with the median background subtracted.
Log Ratio	log (base 2) transform of the ratio of the medians.
Flags	the type of flag associated with a feature.
Normalize	the normalization status of the feature (included/not included).

Example: Affymetrix Rat230_2 platform

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- 31099 probe sets
- 54 control genes
  - Poly-A controls
    - thr, trp, B.suptilis, lys, phe, thr, dap
    - to control target labeling
  - Hybridization control
    - bioP, bioC, bioD, cre
    - to evaluate sample hybridization
  - internal controls
    - β-actin, GADPH, HExokinase
    - to asses the RNA sample and assay quality
- 100 Normalization controls
  - genes that should not vary $\Longrightarrow$ can be used to normalize the measurement
- more details (one the web):
  GeneChip Expression Analysis: Data Analysis Fundamentals

Hippocampus response to bicuculline

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- experiments carried out in Vincent Torre's Lab

Exp. v.s. time	0 min	180 min
exp1	C_{exp1}	T_{exp1}
exp2	C_{exp2}	T_{exp2}
exp3	C_{exp3}	T_{exp3}

- TASK: look for differentially expressed genes

- look at the data preprocessed with MAS 5.0
- genes with "good" P-value in *all* experiments are $\simeq 34\%$ $\Longrightarrow \simeq 10000$ genes
- genes with "good" P-value in *at least 3 out of 6* experiments are $\simeq 51\% \Longrightarrow \simeq 16000$ genes

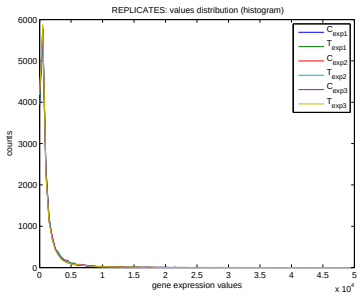
Replicates: histogram

Introduction

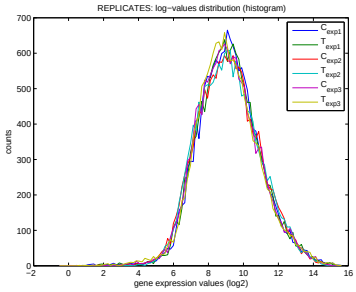
Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- distribution of values: histogram



- take the log of the histogram



distribution are almost normal $\Longrightarrow$"log" is good

Replicates: experiments

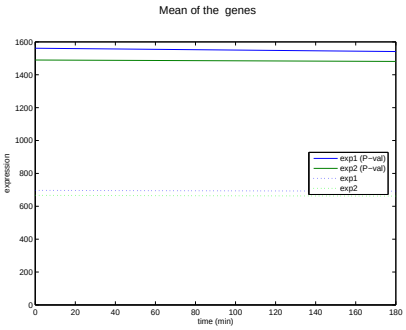
Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- some statistics:

	$C_{\exp 1}$	$T_{\exp 1}$	$C_{\exp 2}$	$T_{\exp 2}$	$C_{\exp 3}$	$T_{\exp 3}$
mean (all):	696.6	691.8	666.4	662.0	675.8	700.7
mean (P-val):	1561.2	1541.1	1489.8	1480.9	1197.3	1235.7
st. dev.(all):	1884	1838	1603	1555	1649	1902
st. dev.(P-val):	2748	2676	2287	2206	2157	2521



- genes with bad P-value have mostly low expression level
- (global) normalization is done over the mean of all data (at least for exp1 and exp2)

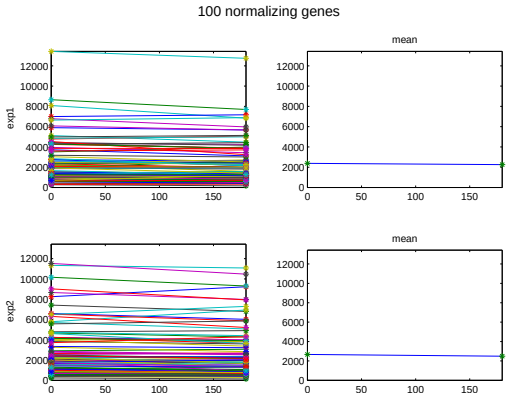
Replicates: 100 normalizing genes

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- looking at the 100 normalization genes



  - they mostly remain constant in each set of experiments (not between the two sets)
  - averages are quite similar (good news)

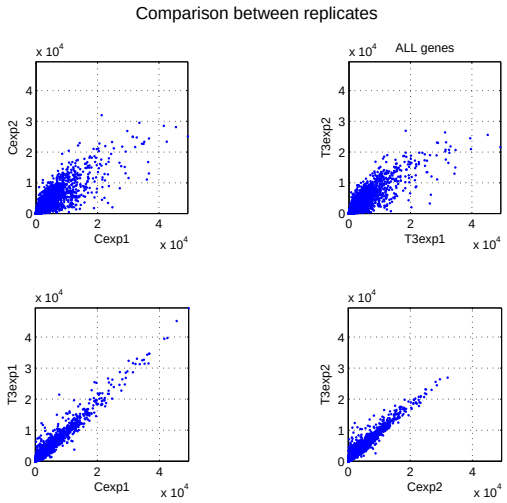
Replicates: scatter plot

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- scatter plot (plot one experiment versus another one)



- no good news!!!!
- variance $\{C_{\text{exp1}} - C_{\text{exp2}}\}$ >> variance $\{C_{\text{exp1}} - T_{\text{exp1}}\}$ and likewise for the others

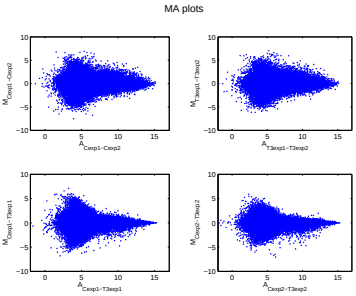
Replicates: systematic error compensation

Introduction

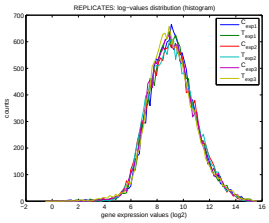
Low level analysis of microarrays

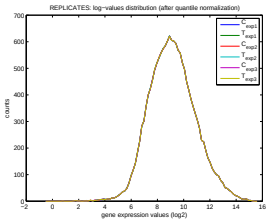
Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- systematic error:
  - biological diversity between sets of experiments
  - also different processing algorithms?
- How to compensate for this systematic error?
- INTERSAMPLE NORMALIZATION
  - check if there is an intensity-dependent bias (MA plot)



  - apply a quantile normalization





Fold change analysis

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- which genes are differentially expressed between C and T?

$$\frac{x_T}{x_C} > k \qquad \text{k-fold up-regulation}$$

$$\frac{x_T}{x_C} < \frac{1}{k} \qquad \text{k-fold down-regulation}$$

- taking log $\Longrightarrow$fold change becomes an additive operation

$$\log\left(\frac{x_T}{x_C}\right) > \log(k) \qquad \Longleftrightarrow \qquad \log(x_T) - \log(x_C) > \log(k)$$

$\Longrightarrow$useful for replicates

Fold change analysis

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot

Fold change analysis
- t-test
- other statistical tests
- multiple testing

- Example: what is the mean fold change in a replicated experiments?

	exp1	exp2
geneA	120	30
geneB	60	60

	exp1	exp2
$\frac{\text{geneA}}{\text{geneB}}$	$\frac{120}{60} = 2$	$\frac{30}{60} = 0.5$

  - mean $\left(\frac{\text{geneA}}{\text{geneB}}\right) = \frac{2+0.5}{2} = 1.25$??????
  - if we take logs:

	exp1	exp2
$\log_2\left(\frac{\text{geneA}}{\text{geneB}}\right)$	$\log_2\left(\frac{120}{60}\right) = 1$	$\log_2\left(\frac{30}{60}\right) = -1$

  - mean $\left(\log_2\left(\frac{\text{geneA}}{\text{geneB}}\right)\right) = \frac{1-1}{2} = 0$,
  - hence $2^0 = 1$

Fold change analysis

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot

Fold change analysis
- t-test
- other statistical tests
- multiple testing

Fold change methods:

- check differentially expressed genes on each experiment pair and take intersection
- previous method: $2^{\text{mean}(\log(T/C))}$
- average over all C and T and then take ratio mean(T)/mean(C)
- example: rat hippocampus

	exp1	exp2	exp3	intersect(1,2,3)	intersect(1,2)
$x_T/x_C > 4$	145	125	62	0	41
$x_T/x_C < 1/4$	72	26	69	0	1

	$2^{\text{mean(log(T/C))}}$	mean(T)/mean(C)	common
$x_T/x_C > 4$	27	26	19
$x_T/x_C < 1/4$	3	0	0

Statistical tests for differential expression

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- But what is a *significant* change???
- Depends on the variability within groups which may be different from gene to gene
- to assess statistical significance of differences: statistical tests for each gene
- two-sample t-statistics

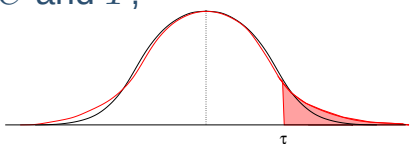
t-test

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- want to compare two groups (here C and T)
- normally distributed, but few samples
  $\Longrightarrow$statistically significant test is t-test
- hypoteses: $\begin{cases} H_0: & \bar{x}_T - \bar{x}_C < \Delta \\ H_1: & \bar{x}_T - \bar{x}_C > \Delta \end{cases}$
- test statistics
- if variances are equal: student t-test $\tau = \frac{(\bar{x}_T - \bar{x}_C) - \Delta}{s\sqrt{\frac{1}{n_C} + \frac{1}{n_T}}}$
  - $s = \sqrt{\frac{(n_C-1)s_C^2 + (n_T-1)s_T^2}{n_C+n_T-2}}$
  - $n_C,\ n_T$ = number of repeats for C and T;
  - $s_C,\ s_T$ = standard deviations
- P-value: area $P(t > \tau)$
  for the t-student curve
  (with $n_c + n_T - 2$ degrees of freedom)
  = probability that the test statistics is at least as extreme as the observed value τ



t-test

Introduction
Low level analysis of microarrays
Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- if variances can be different: Welsh t-test $\tau = \frac{(\bar{x}_T - \bar{x}_C) - \Delta}{\sqrt{\frac{s_C^2}{n_C} + \frac{s_T^2}{n_T}}}$
- n. of degrees of freedom $\nu = \frac{\left(\frac{s_C^2}{n_C} + \frac{s_T^2}{n_T}\right)}{\frac{1}{n_C - 1}\left(\frac{s_C^2}{n_C}\right)^2 + \frac{1}{n_T - 1}\left(\frac{s_T^2}{n_T}\right)^2}$
- P-value: area $P(t > \tau)$ for the t-student curve (with ν degrees of freedom)
- meaning: for each gene weight the differences between C and T by the sample variance of the measures of C and T
- Example: rat hippocampus, overlap with fold change analysis

	t-test (P-val=0.1)	$2^{\text{mean(log(T/C))}}$	mean(T) / mean(C)	common
$x_T/x_C > 4$	0 (0)	27	26	0
$x_T/x_C > 2$	3 (13)	218	254	3 (13)

Alternative statistical tests

Introduction
Low level analysis of microarrays
Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- two-sided tests, $\Delta = 0$,

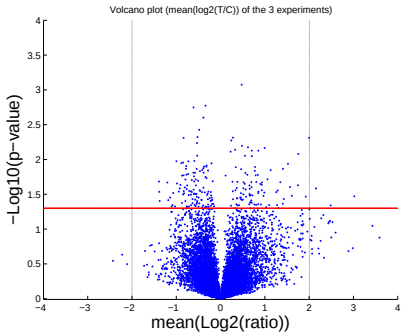
$$\tau = \frac{\bar{x}_T - \bar{x}_C}{s\sqrt{\frac{1}{n_C} + \frac{1}{n_T}}}$$

- Wilcoxon tests: rank-based (non-parametric)
- permutation based tests: in the test statistics choose the null distribution by repeated permutations on the values
- Volcano plots: Log(p-val) vs Log(ratio)
  - abscissa:

$$\log_2\left(\frac{x}{y}\right)$$

  - ordinata

$$-\log_{10}(\text{P-value})$$



Moderate t-statistics

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- with few replicates (2-5 per group) variance estamates are unstable
- in a moderated t-statistics the estimated gene specific variance s is augmented with s_o, a global variance estimator from all genes

$$\tau = \frac{\bar{x}_T - \bar{x}_C}{\sqrt{\nu s + \lambda s_o}\sqrt{\frac{1}{n_C} + \frac{1}{n_T}}}$$

- meaning: “interpolation” between t-statistics and fold-change analysis

Multiple hypotesis testing

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

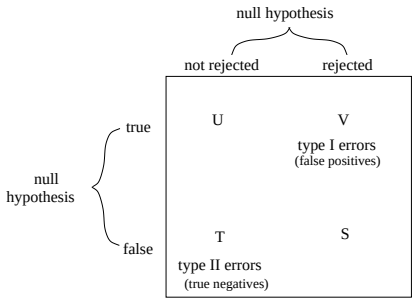
- Problem: testing thousands of hypothesis (one for each gene) simultaneously the chances of false positives increase
- Example: if 3% of 10000 null hypoteses are rejected with a significance value of 0.05 (i.e. 300 genes are differentially expressed)
  $\Longrightarrow$P(single correct rejection) = 1-0.05 =0.95
  $\Longrightarrow$P(correct rejection on 300 genes) = $0.95^{300} = 2.07\ 10^{-7} \simeq 0$
  $\Longrightarrow$P(at least a false rejection) = $1 - 2.07\,10^{-7} \simeq 1$
- Example: if 0% of 10000 genes are differentially expressed at a significance value of 0.01
  $\Longrightarrow$you expect $10000 \cdot 0.01 = 100$ genes to be differentially expressed

Multiple hypotesis testing

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing



- Type I errors = false positives (rejected true null hyp.)
- Type II errors = true negative (not rejected false null hyp.)

- Type I error rates:
  - FWER (Family-Wise Error Rate) = $P(V > 0)$
  - FDR (False Discovery Rates) $= E(Q)$
    with $Q = V/\#$rejections

Multiple hypotesis testing

Introduction

Low level analysis of microarrays

Differentially expressed genes
- Example: Rat230
- Example: Hippocampus
- histogram
- experiments
- 100 normalizing genes
- scatter plot
- Fold change analysis
- t-test
- other statistical tests
- multiple testing

- to control FWER: Bonferroni correction
  - given n genes
  - a test statistics τ
  - an unadjusted P-value p
  - $\Longrightarrow$adjusted P-value = $\min(1, n \cdot p)$
  - very conservative
- to control FDR:
  - order unadjusted P-values $p_{g_1}, p_{g_2}, \ldots, p_{g_n}$
  - to control FDR at a level α

$$j^* = max\{j : p_{g_j} \leqslant \frac{j}{n}\alpha\}$$

  - reject H_0 for $j = 1, \ldots, j^*$
  - conservative if many genes are differentially expressed