BioConductor for the Analysis of Affymetrix Microarray Data

Ann Hess
Department of Statistics
hess@stat.colostate.edu

BioConductor

- "Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data." (www.bioconductor.org)
- BioConductor is based on the R programming language.
- Features of BioConductor:
  - New methods quickly incorporated
  - Flexibility of R programming
  - Help and documentation (help, vignettes, mail list)
  - Free!

Installing R and BioConductor

- To install R go to: www.r-project.org
  Choose a CRAN Mirror
  Download R2.1.1
  (Available for Linux, Mac or Windows)

- To install BioConductor, open R and type:

```
source("http://www.bioconductor.org/biocLite.R")
biocLite()
```

  For more information go to:
  http://www.bioconductor.org/download

Workshop Objectives

- Demonstrate BioConductor packages available for Affymetrix arrays
- Provide an introduction to R/BioConductor
- Demonstrate an approach to sample size determination
- Availability of AEP bioinformatics consulting service

Overview

1. Introduction: Estrogen Case Study
2. Data Visualization and Exploratory Analysis
   - BioConductor Demonstration 1
3. Data Preprocessing and Summarization
4. Identification of Differentially Expressed Genes
   - BioConductor Demonstration 2
4. Following up on Differentially Expressed Genes
   - BioConductor Demonstration 3
6. Simulation Framework for Sample Size Determination

1. Introduction: Estrogen Case Study

- During this workshop we will use publicly available data from a set of eight Affymetrix chips from an experiment designed to measure changes in gene expression in a breast-cancer cell line due to the presence (or absence) of estrogen and due to a time effect (10 hours or 48 hours).

  Scholtens et al (2004) "Analyzing factorial designed microarray experiments", *Journal of Multivariate Analysis* 90, 19-43.

- The data is described in detail in the "Estrogen 2x2 Factorial Design" BioConductor vignette by Denise Scholtens and Robert Gentleman:

  "The investigators in this experiment were interested in the effect of estrogen on the genes in ER+ breast cancer cells over time. After serum starvation of all eight samples, they exposed four samples to estrogen, and then measured mRNA transcript abundance after 10 hours for two samples and 48 hours for the other two. They left the remaining four samples untreated, and measured mRNA transcript abundance at 10 hours for two samples, and 48 hours for the other two."
- mRNA samples were obtained under each of the 4 experimental conditions, reverse transcribed to cDNA, fluorescently labeled, fragmented to cRNA and hybridized to Affymetrix Gene Chip arrays.
- Genes of interest are represented on the arrays using probe sets.

Affymetrix GeneChip® Array

Image courtesy of Affymetrix.

• A scanner is used to create an image of the array.

• The intensity of each “spot” indicates how much binding has occurred at that spot.

• A CEL file contains processed intensity values for each spot (after combining pixel level information).

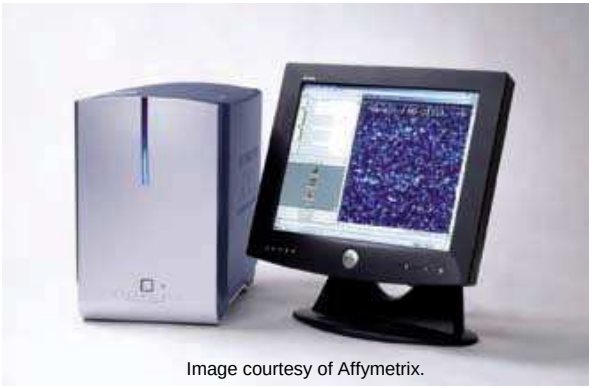


Image courtesy of Affymetrix.

Image of a hybridized Affymetrix GeneChip

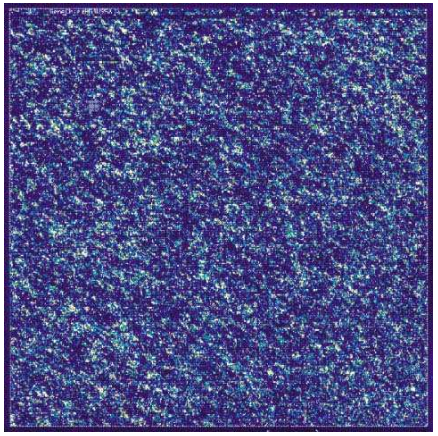


Image courtesy of Affymetrix.

- A CEL file is a quantitative summary of the scanned image.
- For each “spot” location the mean and standard deviation of the pixel level information is given.
- For the HG_U95A arrays used in the estrogen experiment, there are a total of 640x640=409,600 spots on the array.

X	Y	MEAN	STDV	NPIXELS
0	0	107	86.9	25
1	0	7627.5	781.8	20
2	0	135	30.8	25
3	0	7459	1628.2	25
4	0	50	11	20
5	0	97	41.2	25
6	0	7103	1711.4	25
7	0	95.5	14.8	16
8	0	6829	1114.2	20

CEL and CDF files

- CEL files are considered “raw” data.
- To do anything with the CEL file, the CDF (Chip Description File) is also needed.
- The CDF specifies the probe and probe set to which each cell (or spot) belongs.
- CDF information is provided by BioConductor CDF packages which are available for most Affymetrix chip types.
- In most cases, after reading in the CEL files, BioConductor will load the appropriate CDF package (or go to the web to find it).

Gene expression → Spot Intensity

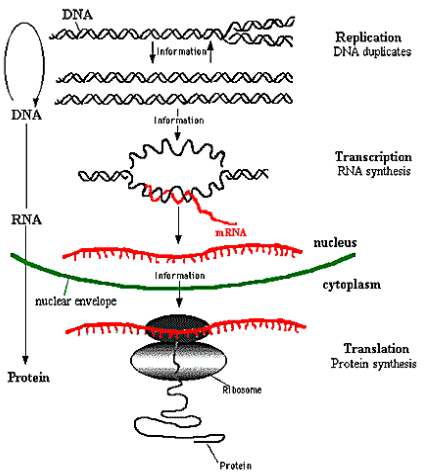
- Genes express themselves by producing mRNA.
- The more active a gene is the more mRNA it will produce.
- mRNA is used to make proteins which are responsible for cell function.
- The intensity of a spot is indicative of how much labeled transcript has bound to that spot.
- Brighter spots indicate higher transcript abundance and therefore more active genes.

Central Dogma of Molecular Biology

DNA contains genetic information.

mRNA carries information from DNA and serve as templates that order different proteins.

Proteins are responsible for cell function.



The Central Dogma of Molecular Biology

Image courtesy of Access Excellence @ the National Health Museum

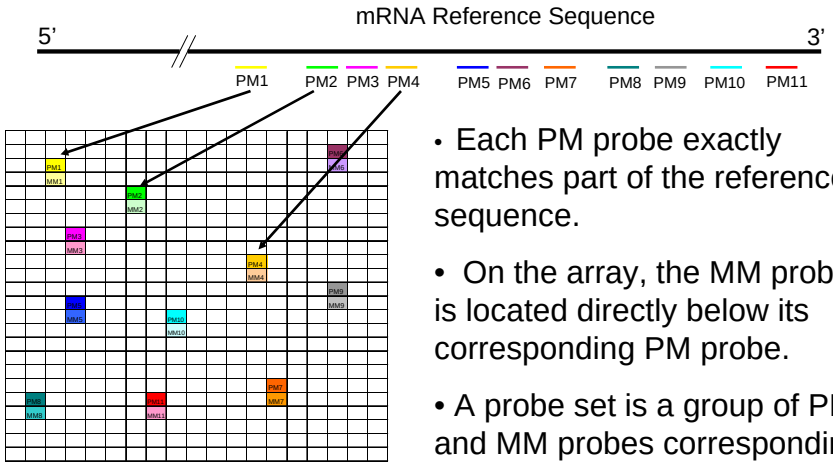
Probe Level Information

- Short sequences of DNA (called oligonucleotides or oligos) represent each gene.
- Each probe is a 25 base sequence.
- A probe set is a group of probes that corresponds to a particular gene or EST (expressed sequence tag).
- Most genes or ESTs are represented by a single probe set, but some are represented by more than one probe set.

Perfect Match (PM) and MisMatch (MM) probes

- A probe pair consists of a
  - PM (perfect match) probe which exactly represents part of the DNA sequence for a gene or EST.
  - MM (mismatch) probe which differs from the PM probe only at the middle base (A↔T, C↔G).
- Intensity readings from PM probes represent gene specific binding (GSB) plus some binding due to cross hybridization (nonspecific binding).
- Intensity readings from MM probes can be used to account for nonspecific binding (NSB).
- For the estrogen experiment, 12625 probe sets (typically with 16 probe pairs) are present on each HG_U95A array.

Probes and Sequences

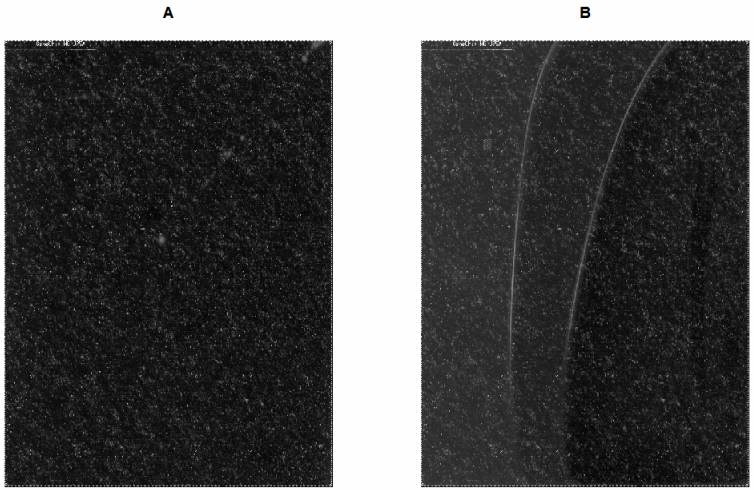


- Each PM probe exactly matches part of the reference sequence.
- On the array, the MM probe is located directly below its corresponding PM probe.
- A probe set is a group of PM and MM probes corresponding to a gene or EST.

2. Data Visualization and Exploratory Analysis

- Analysis starts with CEL files as the raw data.
- Before we attempt to identify "differentially expressed genes" we need to check data quality.
- As an example, we will look at a group of Affymetrix HGU95A GeneChips labeled A-F.
- This data is available from the jsmHyperdip BioConductor package and is used here for illustration purposes only.

CEL file Images

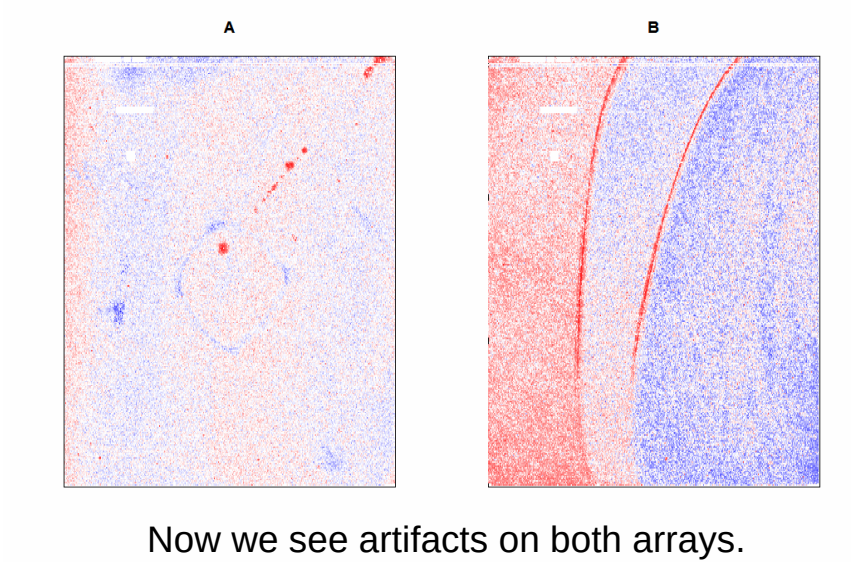


Array B has an artifact.

Residual Images using affyPLM

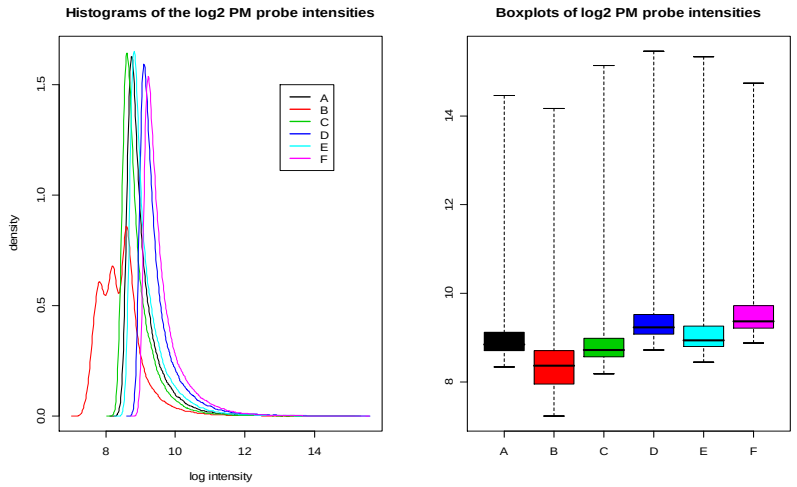
- affyPLM fits a probe set level linear model.
- By (graphically) examining the residuals, affyPLM offers another way to look for array artifacts.
- The residual images are colored so that large positive residuals are red, large negative residuals are blue and small residuals are white.

Residual Images



Now we see artifacts on both arrays.

Histograms and Box plots of the probe intensities by Array.

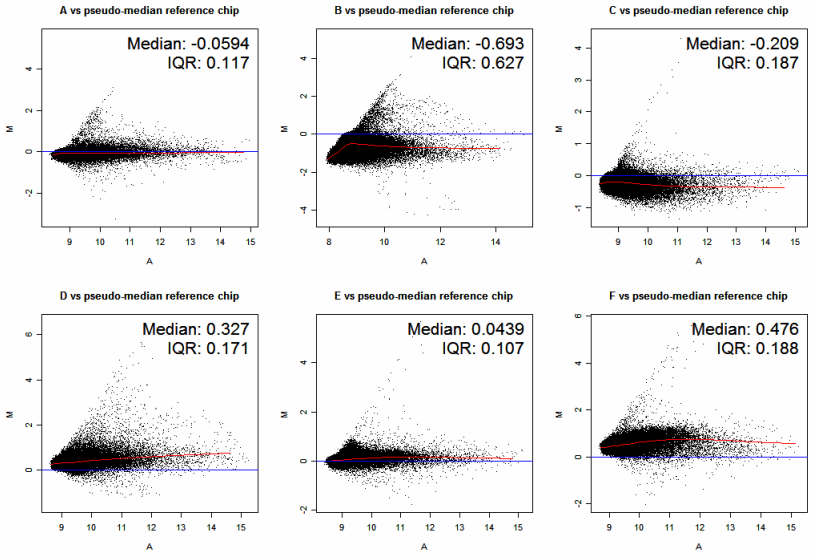


The histogram for Array B has a distinctly different shape.

MA plots

- M values are log fold changes.
  M=log2(T/C)=log2(T)-log2(C),
  where T represents a value from a treatment array and C represents a value from a control array.
- A values are average log intensities between two arrays.
  A=(log2(T)+log2(C))/2
- Since we assume the majority of genes will not be differentially expressed, we would like the observations to scatter around M=0.
- *After* normalization, this should be the case.

MA plots



Demonstration 1

- We will run some exploratory analyses on the estrogen data:
  - CEL file images
  - Residual images
  - Histograms and box plots
  - MA plots
  - PM/MM correlation

Experimental Design for the Estrogen Experiment

Time	Estrogen	
	Absent	Present
10 hours	low10-1.cel low10-2.cel	high10-1.cel high10-2.cel
48 hours	low48-1.cel low48-2.cel	high48-1.cel high48-2.cel

This information is summarized in the file EstrogenPData.txt.

3. Preprocessing and Summarization

- A background correction is usually performed to adjust for optical noise (intensity not related to hybridization).
- Sometimes a nonspecific binding (NSB) correction is performed as part of the background correction.
- "Normalization" is used to correct for systematic array differences.
- Probe set summaries are calculated.

Popular Methods

- Affymetrix Microarray Suite (MAS) 5.0 or Gene Chip Operating System (GCOS) 1.0
- Model Based Expression Index (MBEI) implemented through dChip (www.dchip.org)

  Li and Wong (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *PNAS* 98, 31-36.
- Robust Multichip Analysis (RMA)

  Irizarry et al. (2003) Exploration, Normalization and Summaries of High Density Oligonucleotide Array Probe Level Data, *Biostatistics* 4, 249-264.
- GCRMA

  Wu et al (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays, *JASA* 99, 909-917.

MAS 5.0

Background correction:

$E_{ij} = \log_2(PM_{ij} - IM_{ij})$ fixed probe set, array i and probe *j,*

where IM (Ideal Mismatch) is chosen so IM<PM.

Normalization:

Scale so that the trimmed mean of the E_{ij} values is the same for each chip.

Probe set summary:

$\log_2(S_i) = \text{TukeyBiweight}(E_{i1},...,E_{im})$

The Tukey biweight algorithm is a method to determine a robust average unaffected by outliers.

dChip/MBEI

Background correction:

Can use $E_{ij} = PM_{ij} - MM_{ij}$ for fixed ps, array *i* and probe *j.*

For the PM only model, optical noise correction is performed on PM values.

Invariant Set Normalization:

A normalization curve is fit to a group of PM probes which are thought to be unchanged.

Probe set summary (MBEI):

Fit $N(E_{ij}) = N(PM_{ij}) - N(MM_{ij}) = \theta_i \varphi_j + \varepsilon_{ij}$

or $N(PM_{ij}) = \nu_i + \theta_i \varphi_j + \varepsilon_{ij}$

Probe Set Summary = θ_i

RMA

NOTE: RMA uses PM values only in all steps of analysis!

Background correction:

$E_{ij} = PM_{ij} - B_{ij}$ for fixed probe set, array *i* and probe *j,*

where B is estimated so B<PM.

Quantiles Normalization:

Forces distribution of PM values to be the same for every array in an experiment.

Probe set summary (RMA):

Robustly fit $\log_2 N(E_{ij}) = \mu_i + \alpha_j + \varepsilon_{ij}$

Probe Set Summary = μ_i

GCRMA

Background correction:

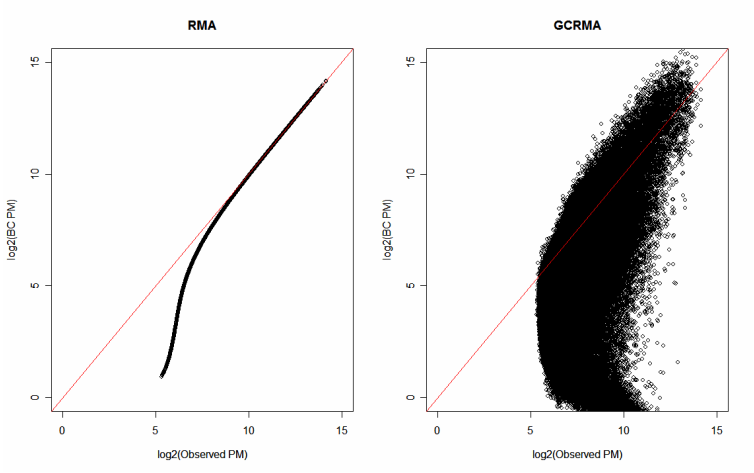
Wu *et al.* note that “RMA does not adjust well for nonspecific binding.”

They incorporate probe sequence to better estimate NSB.

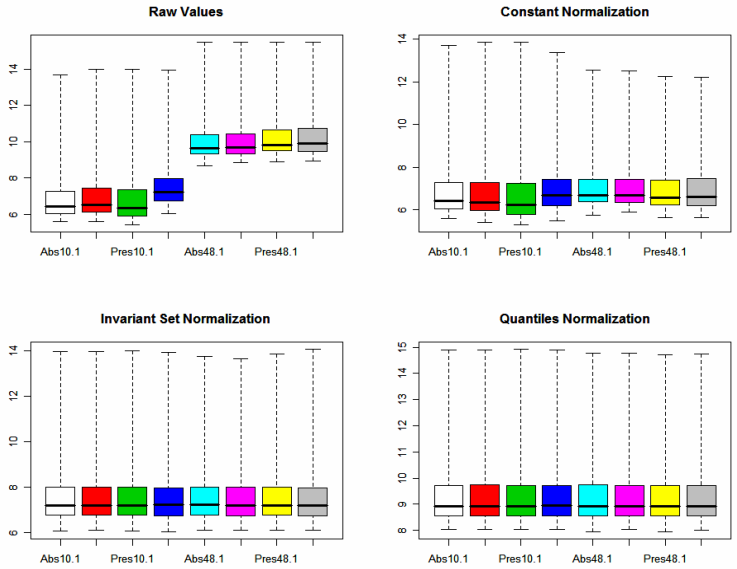
Quantiles Normalization

Probe set summary (RMA)

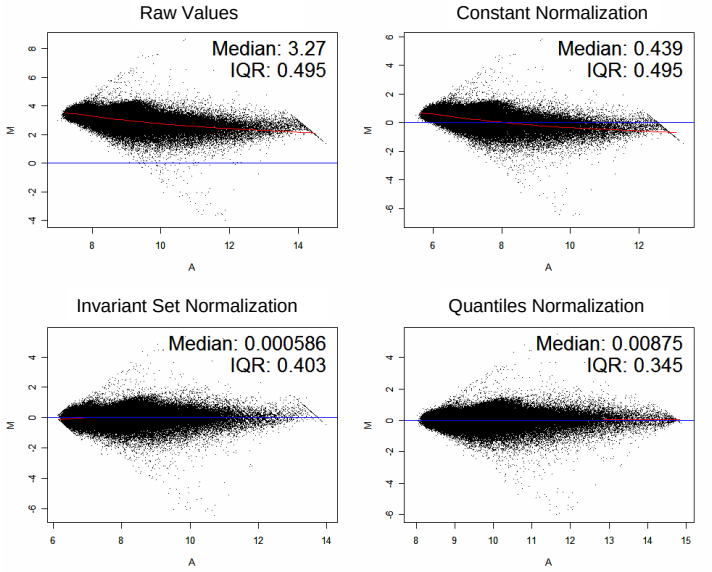
Background Corrected PM values versus Observed PM values



Boxplots of Observed and Normalized Data



MA plots of Observed and Normalized Data



Comparison of Methods

- What is the “best” method? How do the methods perform?
- If we have data where the truth is known, we can compare the performance of the methods.
- The “Golden Spike” data has

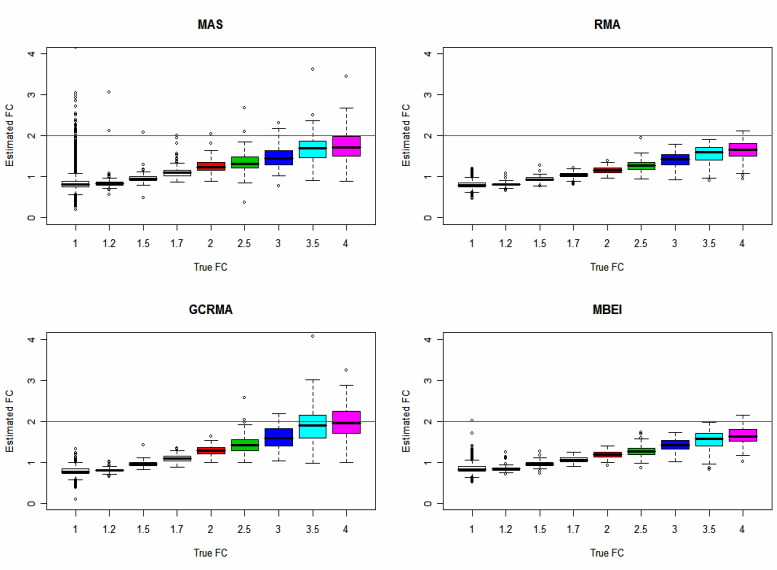
  - 1331 probe sets “spiked-in” at known fold changes between 1.2 and 4,

  - 2535 probsets spiked in with known fold change of 1,

  - 10144 empty probe sets.

  Choe et al. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset, *Genome Biology* 6:R16
- Used drosgenome1 Affymetrix GeneChips.

Boxplots of Estimated FC by Method



Power and FDR

- Power is defined as the probability that a test will declare a gene to be differentially expressed when in fact this is true.
- To estimate power, we consider the proportion of genes that were declared differentially expressed, when the true FC is less than 2 or greater than 2.
- False discovery rate (FDR) is estimated as the proportion of genes incorrectly declared to be differentially expressed.
- The p-values were corrected (for multiple comparisons) using the Benjamini-Hochberg (1995) method and a p-value cut off of 0.05 was used.
- The moderated t-statistic was used to test for differential expression.

Observed Power and FDR by Method

	MAS	RMA	GCRMA	MBEI
FDR	0.69	0.79	0.69	0.79
Power (FC<2)	0.23	0.46	0.39	0.51
Power (FC≥2)	0.79	0.89	0.90	0.90

We see that the FDR is much higher than expected!

However, if we consider the 100 probe sets with the smallest (adjusted) p-values, we find that:

72% are truly differentially expressed for MAS,

96% are truly differentially expressed for RMA,

90% are truly differentially expressed for gcRMA,

85% are truly differentially expressed for MBEI.

MAS Present/Absent Calls

- MAS uses PM and MM information to test whether a probe set is present (signal greater than background).
- For the “Golden Spike” data, we know which probe sets are empty.
- We can check the accuracy of the MAS Present/Absent Calls (as implemented in BioC):

  ~88% of nonempty probe sets were called “Present”.

  ~92% of empty probe sets were called “Absent”.

4. Identification of Differentially Expressed Genes

- Testing for differential expression starts with probe set summary values (expression values).

← Control Arrays → ← Treatment Arrays →

genes	Array 1	Array 2	Array 3	Array 4	Array 5	Array 6
	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}
	s_{21}	s_{22}	s_{23}	s_{24}	s_{25}	s_{26}
	…	…	…	…	…	…
	s_{n1}	s_{n2}	s_{n3}	s_{n4}	s_{n5}	s_{n6}

- To test for differential expression we consider:

  H_0: $\mu_C=\mu_T$ (not differentially expressed)

  versus H_a: $\mu_C\neq\mu_T$ (differentially expressed)

  We perform a test for every gene!

Test Statistics

- t-statistic: $t_g = \dfrac{\overline{M}_g}{SE_g}$

$$\overline{M}_g = \log_2(FC_g) = \log_2(\overline{T}_g / \overline{C}_g) = \log_2(\overline{T}_g) - \log_2(\overline{C}_g)$$

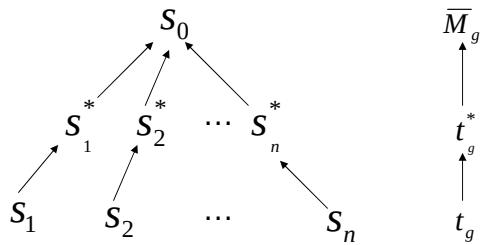
SE_g is the standard error for gene *g*.

Genes with small sample variance are more likely to be called differentially expressed.

- Moderated t statistic: $t_g^* = \dfrac{\overline{M}_g}{SE_g^*}$

Where SE* is calculated after the sample variance is "shrunk" towards some common value.

Shrinkage of standard deviations



The data decide whether t_g^* should be closer to $\overline{M}_g$ or t_g.

The moderated t-statistic reduces the impact of large and small sample standard deviations.

Why use the Moderated t-statistic?

- According to research by Smyth, the moderated t-statistic has lower FDR and higher power than competing methods (ordinary t-statistic, Efron's empirical Bayes method).

  Smyth (2004) "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments" *Statistical Applications in Genetics and Molecular Biology* 3(1), article 3.
- Flexibility to handle contrasts.
- Easy to implement in Bioconductor!

Linear Models

- Fit a linear model for each gene g:
$$E(Y_g) = X\beta_g$$
where X is the design matrix, Y_g is the vector of expression values and β_g represents the estimated coefficients, estimated by $\hat{\beta}_g$.
- Linear models allows us to combine information across arrays. They can handle arbitrarily complicated experiments.
- REQUIRED:
Design Matrix specifies the experimental design.
Contrast Matrix specifies which comparisons are of interest.

Contrasts

- A contrast is a combination of population means of the form
$$\Psi = \sum a_i \mu_i.$$
where the coefficients a_i have sum zero.
- The corresponding sample contrast is $c = \sum a_i \bar{x}_i$.
- The standard error of c is $SE_c = s_p \sqrt{\sum \frac{a_i^2}{n_i}}$.
- To test the null hypothesis $H_0 : \Psi = 0$ use the t statistic $t = \frac{c}{SE_c}$.

Contrasts Matrix

- Contrasts are linear combinations of parameters from the linear model fit.
$$\hat{\alpha}_g = C^T \hat{\beta}_g$$
where $\hat{\alpha}_g$ is a vector of contrasts for gene *g*, *C* is the contrasts matrix, and $\hat{\beta}_g$ is a vector of coefficients from a linear model fit.
- The columns of C correspond to contrasts.
- The rows of C correspond to treatments.

Design Matrix for the Estrogen Data

- For the estrogen data, we have 8 arrays and 4 treatments (Abs10, Pres10, Abs48, Pres48).
- The Design Matrix:

X=

← treatments →				
Abs10	Pres10	Abs48	Pres48	
1	0	0	0	Abs10.1
1	0	0	0	Abs10.2
0	1	0	0	Pres10.1
0	1	0	0	Pres10.2
0	0	1	0	Abs48.1
0	0	1	0	Abs48.2
0	0	0	1	Pres48.1
0	0	0	1	Pres48.2

arrays

Contrasts Matrix for the Estrogen Data

- For the estrogen experiment, there are three comparisons of interest:

- c1: Pres10 versus Abs10 (estrogen effect at 10 hours)
- c2: Pres48 versus Abs48 (estrogen effect at 48 hours)
- c3: Abs48 versus Abs10 (time effect without estrogen)

$$C = \begin{array}{c} \overset{\longleftarrow \text{contrasts} \longrightarrow}{\begin{matrix} C1 & C2 & C3 \end{matrix}} \\ \begin{pmatrix} -1 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{matrix} Abs10 \\ Pres10 \\ Abs48 \\ Pres48 \end{matrix} \end{array} \quad \begin{matrix} \uparrow \\ \text{treatments} \\ \downarrow \end{matrix}$$

Multiple Testing Adjustment

- Suppose we have an array representing 10,000 genes of which 2% are differentially expressed. We choose α=0.05. With 100% power, we would find:
  - 200 (truly) differentially expressed genes
  - 490 false positives
- False discovery rate (FDR) = 490/(200+490) = 0.71
- The Benjamini Hochberg multiple testing adjustment attempts to control the FDR.
- Other multiple testing adjustments are available.

The limma Package

- The limma package allows us to use linear models to analyze designed microarray experiments.
- Design matrix and contrasts matrix are required.
- limma uses empirical Bayes (moderated t) method.
- Multiple testing adjustments can also be done using limma.

A Note about Ranking

- "In many gene discovery experiments for which microarrays are used, the primary aim is to rank the genes in order of evidence against H_0 rather than assign absolute p-values. This is because only a limited number of genes may be followed up for further study regardless of the number which are significant" Smyth(2004)
- Ranking is easier than testing!

Demonstration 2

- Preprocess the data according to the RMA algorithm and examine the box plots and MA plots.
- Calculate probe set summaries using RMA.
- Identify differentially expressed genes for various contrasts using moderated t-statistic (eBayes).
- Export a table of log fold changes and p-values.

5. Following up on Differentially Expressed Genes

So, we have a list of differentially expressed genes….what do we do now?

- Annotation
- Gene Ontology
- Venn Diagrams
- Clustering
- Diagnostics

Annotation

- Annotation Information is available for many Affymetrix Chip types from BioConductor: www.bioconductor.org/data/metaData.html
- Information includes
  - Probe set ID
  - Gene symbol/name
  - UniGene ID

Gene Ontology

"The Gene Ontology, or GO, is composed of three related ontologies covering basic areas of biological research: the molecular function of gene products, their role in multi-step biological processes, and their physical structure as cellular components. Each ontology is constructed as a directed acyclic graph."

http://en.wikipedia.org/wiki/Gene_ontology

The Ontologies

- F: “Molecular function describes activities, such as catalytic or binding activities, at the molecular level.”
- P: “A biological process is series of events accomplished by one or more ordered assemblies of molecular functions.”
- C: “A cellular component is just that, a component of a cell but with the proviso that it is part of some larger object.”

  http://www.geneontology.org/GO.doc.shtml#ontologies

An Example of GO Results for TFF1

TFF1_HUMAN

Full Name: None
Type: protein
Synonyms: IPI00022283
Datasource: UniProt (The Universal Protein Resource)

Associated to Terms:

Term	Ontology	Evidence	Reference
carbohydrate metabolism	P	TAS	PMID:2303034
defense response	P	NR	UniProt:P04155
digestion	P	NAS	PMID:9043862

Evidence Codes:

NAS: Non-traceable Author Statement

NR: Not Recorded

TAS: Traceable Author Statement

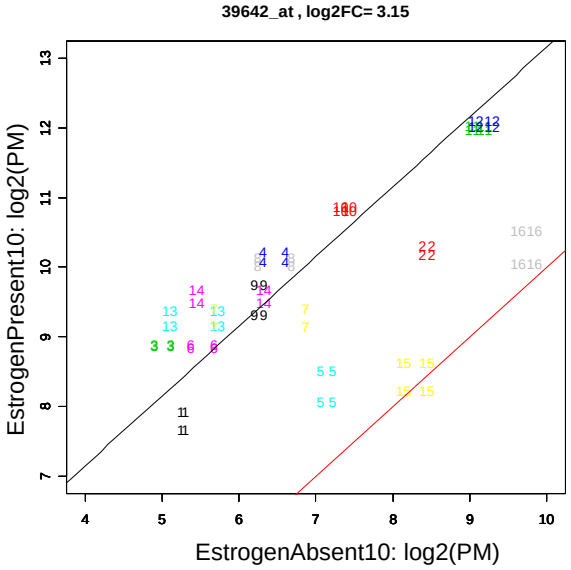
Probe Level Diagnostics

- For a given probe set, each probe can be used to estimate the fold change.

	EstAbs10.1	EstAbs10.2	EstPres10.1	EstPres10.2
probe 1	$s_{1,1}$	$s_{1,2}$	$s_{1,3}$	$s_{1,4}$
probe 2	$s_{2,1}$	$s_{2,2}$	$s_{2,3}$	$s_{2,4}$
….				
probe 16	$s_{16,1}$	$s_{16,2}$	$s_{16,3}$	$s_{16,4}$

- Using only probe *i*, we can get four estimates of fold change: $s_{i,3}/s_{i,1}$, $s_{i,3}/s_{i,2}$, $s_{i,4}/s_{i,1}$, $s_{i,4}/s_{i,2}$.
- Obviously these are NOT independent estimates, but by looking at a plot of the pairwise probe level data, we may be able to spot an outlier probe.

Plot of Probe Values



For each probe, we get four pairs of coordinates ($PM_{Abs10.i}$,$PM_{Pres10.i}$).

The black line indicates the estimated $\log_2$FC. This is estimated as the median of the pairwise probe level $\log_2$FC.

The red line indicates $\log_2$FC=0, for reference.

Clustering

- The goal of clustering is to group observations that are “similar” based on some criteria.
- Clustering can be applied to rows (genes) and/or columns (arrays) of an expression data matrix.
- Clustering allows for reordering of the rows/columns of an expression matrix for easier visualization.
- Clustering is basically an exploratory tool.
- Many clustering methods available in Bioconductor.

Some Clustering Methods

- Hierarchical Clustering: Let each gene be represented by a vector of expression values of length N, where N is the number of arrays. Compute a distance matrix that gives the distances between all pairs (using a distance or similarity measure). Group the two closest genes and define a new node. Continue the process…
- K-means Clustering: User chooses the number of clusters and through an iterative process, each gene is assigned to one of the clusters.
- Self-Organizing Maps: Similar to K-means clustering, but the clusters are arranged in a two-dimensional grid. The grid size is specified by the user.

Demonstration 3

- Annotation
- Gene Ontology
- Probe Level Diagnostic Plot
- Venn Diagrams
- Clustering

6. Simulation Framework for Sample Size Determination

- SimArray is a program for determining sample size by simulation.
- Required input includes at least one “starter” array, estimated fold changes and variance components.
- From this initial input, SimArray simulates microarray data for a requested number of replicates from which the power and FDR can be estimated.

SimArray Algorithm

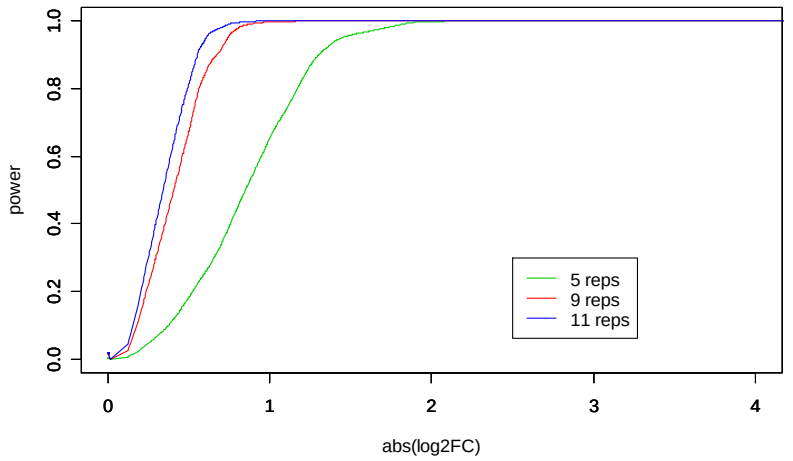
- We start with a model that incorporates the MAS, RMA/GCRMA and MBEI models.
- Fold changes and variance components can be estimated if there is more than one starter array.
- Starter arrays are background corrected and normalized.
- Choose a “true” baseline array from one of the “starter” arrays. Create a “true” experimental array by multiplying the “true” baseline arrays by the assumed fold changes.
- Generate replicates by imposing error using estimated variance components.
- Analyze the simulated data.
- Estimate power and FDR.

Example: Breast Cancer Data

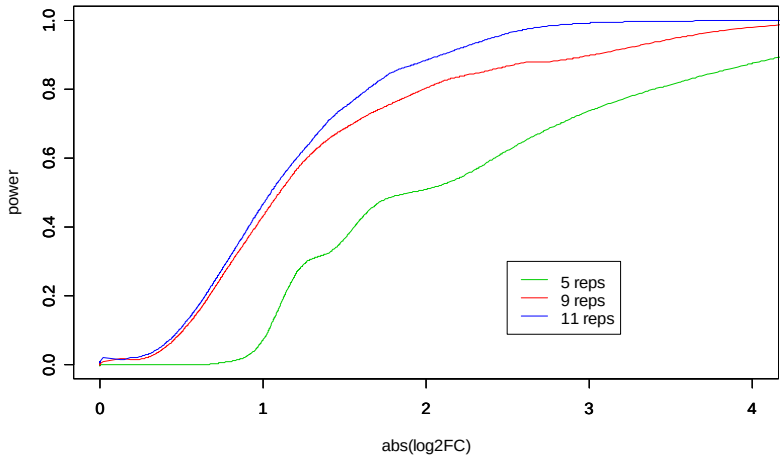
- Dr. Henry Thompson is interested in comparing expression profiles for mice with breast cancer with and without a preventative treatment.
- A pilot study was conducted. Breast cancer was induced in all mice. For each mouse, samples were taken from the tumor and the non-cancerous mammary gland (MG).
- The following pooled samples were obtained and each was represented on a single Affymetrix RAE230A array:
  - untreated tumor
  - untreated MG
  - treated tumor
  - treated MG
- Here we will compare treated tumor versus treated MG.

- A regression model can be fit to each probe set:
  $PM_{T.ij} = b_i PM_{MG.ij} + e_{ij}$ for probe set *i*, probe j.
  where $\hat{b}_i$ is an estimate of fold change.
- We can also test whether b_i = 1.
- When comparing treated tumor versus treated MG, we found that 9% of genes were differentially expressed.
- Variance components were also estimated from the (preprocessed) starter arrays.
- Data was simulated for 5, 9 and 11 replicates per treatment.
- Simulated data was analyzed using RMA and GCRMA.

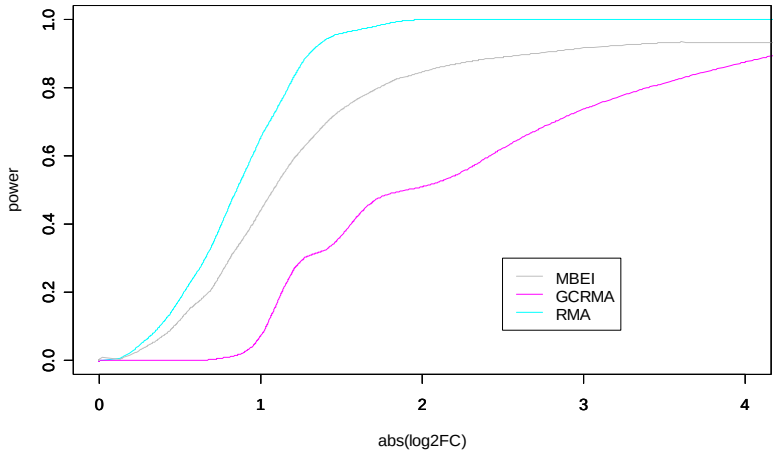
RMA Power



GCRMA Power



Comparison of Methods with 5 Replicates



Estimated FDR

Method	replicates	mean(FDR)	sd(FDR)
MBEI	5	0.079	0.010
gcRMA	5	0.040	0.021
	9	0.163	0.080
	11	0.231	0.095
RMA	5	0.055	0.007
	9	0.158	0.009
	11	0.213	0.010

Discussion of Results

- The power plots shows that substantial gains are made when the number of replicates is increased from 5 to 9. Only modest gains are achieved when the number of replicates is increased from 9 to 11.
- For this data set, RMA appears to have higher power than either GCRMA or MBEI.
- The estimated FDR increases with number of replicates. This may be due to the fact that none of the models accounts for all sources of variation.

Other Considerations when Planning a Microarray Experiment

- Number of arrays
- Types of Samples
  - Replication – technical, biological
  - Pooled versus individual samples
  - Pooled versus amplified samples
- Avoidance of bias
  - experimental conditions, mRNA extraction and processing, the reagents, the operators, the scanners, and so on can leave a "global signature" in the resulting expression data.
  - Randomization!

Some Concluding Remarks about BioConductor

- BioConductor is very flexible.
- MANY options are available:

  Normalization: constant, contrasts, invariant set, loess, qspline, quantiles, VSN

  Testing: Wilcoxon rank sum test, t-statistics, moderated t-statistics, SAM, EBAM

  Multiple Testing Adjustments: Bonferroni, Holm, Hochberg, Sidak, Benjamini-Hochberg, Westfall and Young.

  Clustering: hierarchical, k-means, PAM, SOM
- Good documentation is available.
- Mail list provides prompt response (often from the BioConductor core development team).

Abbreviations

EST: expressed sequence tag
FC: fold change
FDR: false discovery rate
GSB: gene specific binding
MM: mismatch probe
MAS: Microarray Suite
MBEI: model based expression index
NSB: nonspecific binding
PM: perfect match probe
RMA: robust multichip analysis

References

- Benjamini and Hochberg (1995) "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society B*, 57(1), 289-300.
- Choe et al. (2005) "Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset", *Genome Biology* 6:R16.
- Irizarry et al. (2003) "Exploration, Normalization and Summaries of High Density Oligonucleotide Array Probe Level Data", *Biostatistics* 4, 249-264.
- Li and Wong (2001) "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection", *PNAS* 98, 31-36.
- Scholtens et al. (2004) "Analyzing factorial designed microarray experiments", *Journal of Multivariate Analysis* 90, 19-43.
- Smyth (2004) "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments" *Statistical Applications in Genetics and Molecular Biology* 3(1), article 3.
- Wu et al. (2004) "A Model-Based Background Adjustment for Oligonucleotide Expression Arrays", *JASA* 99, 909-917.