Microarray Analysis

The Basics

Thomas Girke

December 9, 2011

Technology

Challenges

Data Analysis

Data Depositories

R and BioConductor

Homework Assignment

Outline

Microarray and Chip Technology

Definition

- Hybridization-based technique that allows simultaneous analysis of thousands of samples on a solid substrate.

Applications

- Transcriptional Profiling
- Gene copy number
- Resequencing
- Genotyping
- Single-nucleotide polymorphism
- DNA-protein interaction (e.g.: ChIP-on-chip)
- Gene discovery (e.g.: Tiling arrays)
- Identification of new cell lines
- Etc.

Related technologies

- Protein arrays
- Compound arrays

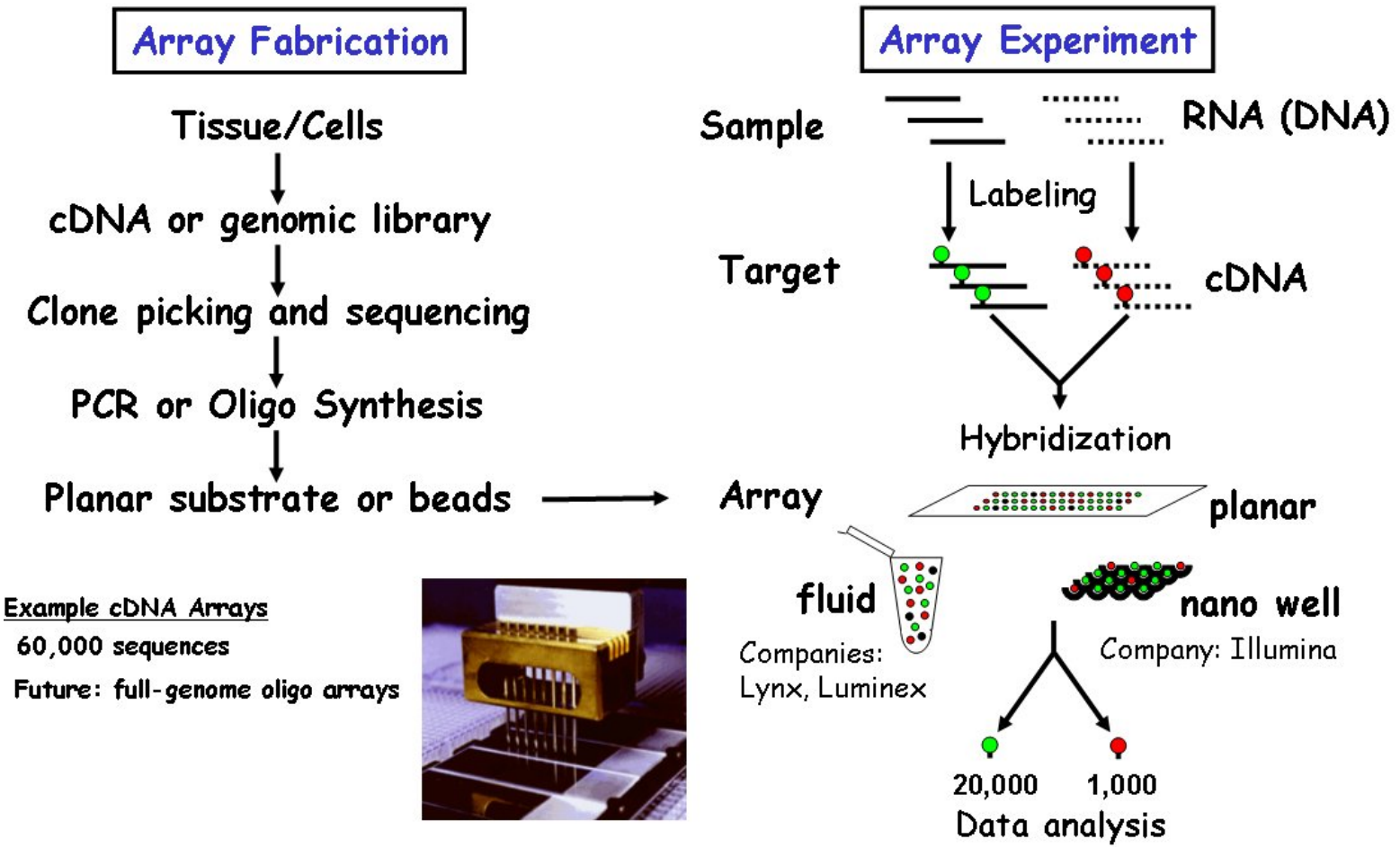
Why Microarrays?

- Simultaneous analysis of thousands of genes
- Discovery of gene functions
- Genome-wide network analysis
- Analysis of mutants and transgenics
- Identification of drug targets
- Causal understanding of diseases
- Clinical studies and field trials

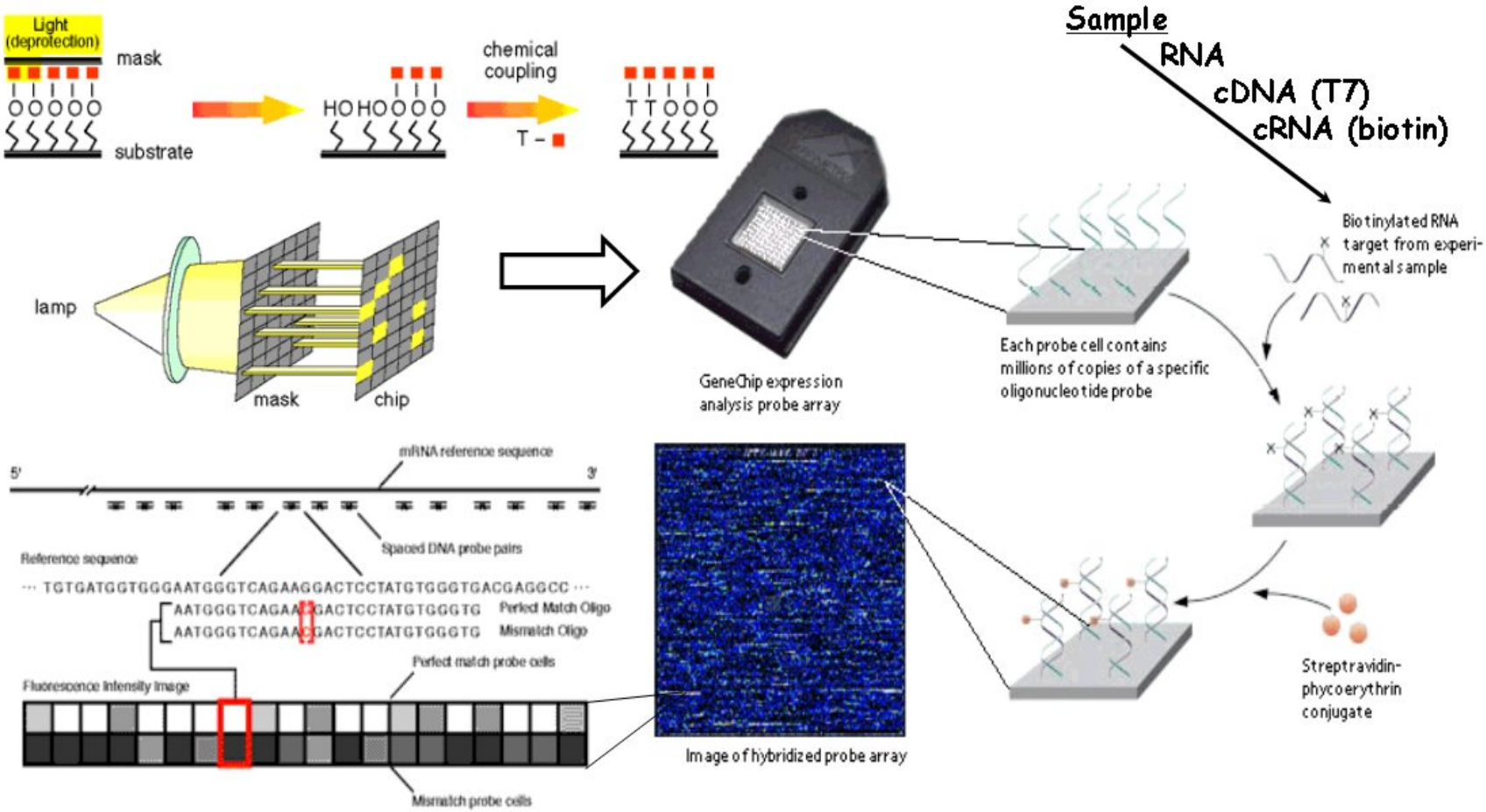
Different Types of Microarrays

- Single channel approaches
  - Affymetrix gene chips
  - Macroarrays
- Multiple channel approaches
  - Dual color (cDNA) microarrays
- Specialty approaches
  - Bead arrays: Lynx, Illumina, ...
  - PCR-based profiling: CuraGen, ...

Dual Color Microarrays

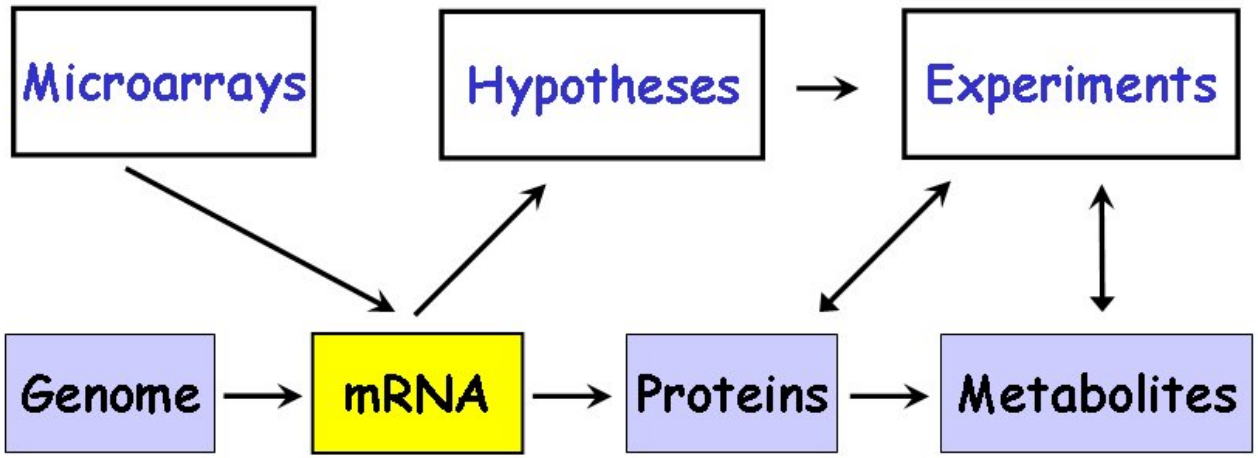


Affymetrix DNA Chips



Outline

Profiling Chips Monitor Differences of mRNA Levels



Efficient strategy for down-stream follow-up experiments important!

Strategies to Validate Array Hits

- Real-time PCR, Northern, etc.
- Transgenic tests
- Knockout plants and/or activation tagged lines
- Protein profiling
- Metabolic profiling
- Other tests: in situ hybs, biochemical and physiological tests
- Integration with sequence, proteomics and metabolic databases

Sources of Variation in Transcriptional Profiling Experiments

- Every step in transcriptional profiling experiments can contribute to the inherent 'noise' of array data.
- Variations in biosamples, RNA quality and target labeling are normally the biggest noise introducing steps in array experiments.
- Careful experimental design and initial calibration experiments can minimize those challenges.

Experimental Design

- Biological questions:
  - Which genes are expressed in a sample?
  - Which genes are differentially expressed (DE) in a treatment, mutant, etc.?
  - Which genes are co-regulated in a series of treatments?
- Selection of best biological samples and reference
  - Comparisons with minimum number of variables
  - Sample selection: maximum number of expressed genes
  - Alternative reference: pooled RNA of all time points (saves chips)
- Develop validation and follow-up strategy for expected expression hits
  - e.g. real-time PCR and analysis of transgenics or mutants
- Choose type of experiment
  - common reference, e.g.: S1 x S1+T1, S1 x S1+T2
  - paired references, e.g.: S1 x S1+T1, S2 x S2+T1
  - loop & pooling designs
  - many other designs
- At least three (two) biological replicates are essential
  - Biological replicates: utilize independently collected biosamples
  - Technical replicates: utilize often the same biosample or RNA pool

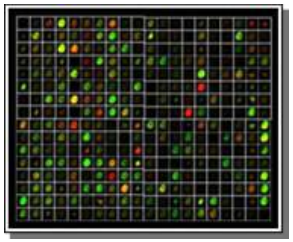
Outline

Basic Data Analysis Steps

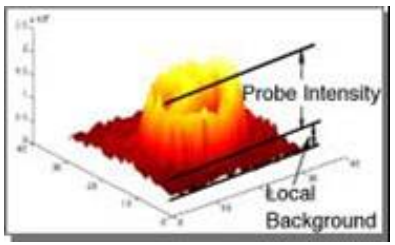
- Image Processing: transform feature and background pixel into intensity values
- Transformations
  - Removal of flagged values (optional)
  - Detection limit (optional)
  - Background subtraction
  - Taking logarithms
- Normalization
- Identify EGs and DEGs
  - Which genes are expressed?
  - Which genes are differentially expressed?
- Cluster analysis (time series)
  - Which genes have similar expression profiles?
- Promoter analysis
- Integration with functional information: pathways, etc.

Image Analysis

- Overall slide quality
- Grid alignment (linkage between spots and feature IDs)



- Signal quantification: mean, median, threshold, etc.



- Local background
- Manual spot flagging
- Export to text file

Image analysis software (selection)

ScanAlyze (http://rana.lbl.gov/EisenSoftware.htm)

TIGR SpotFinder (http://www.tigr.org/software/)

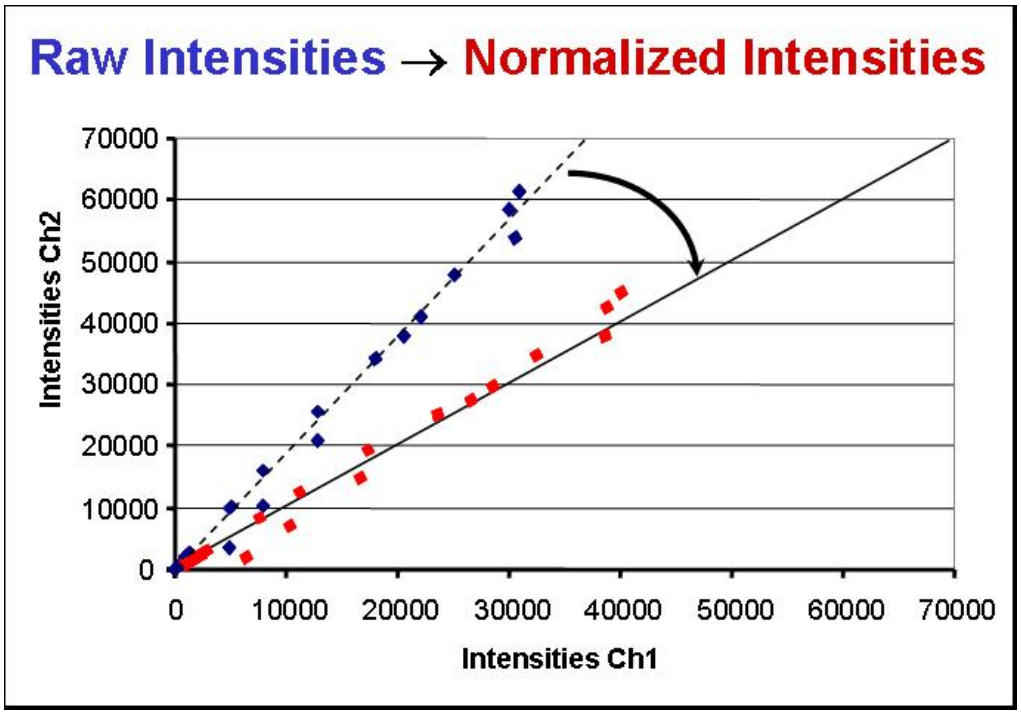
Background Correction

- Filtering (optional)
  - Intensities below detection limit
  - Negative intensities
  - Spacial quality issues
- Background correction
  - BG consists of non-specific hybridization and background fluorescence
  - If BG is higher than signal: (1) remove values, (2) set signal to lowest measured intensity, (3) many other approaches
  - BG subtraction
    - Local background
    - Global background
    - No background subtraction

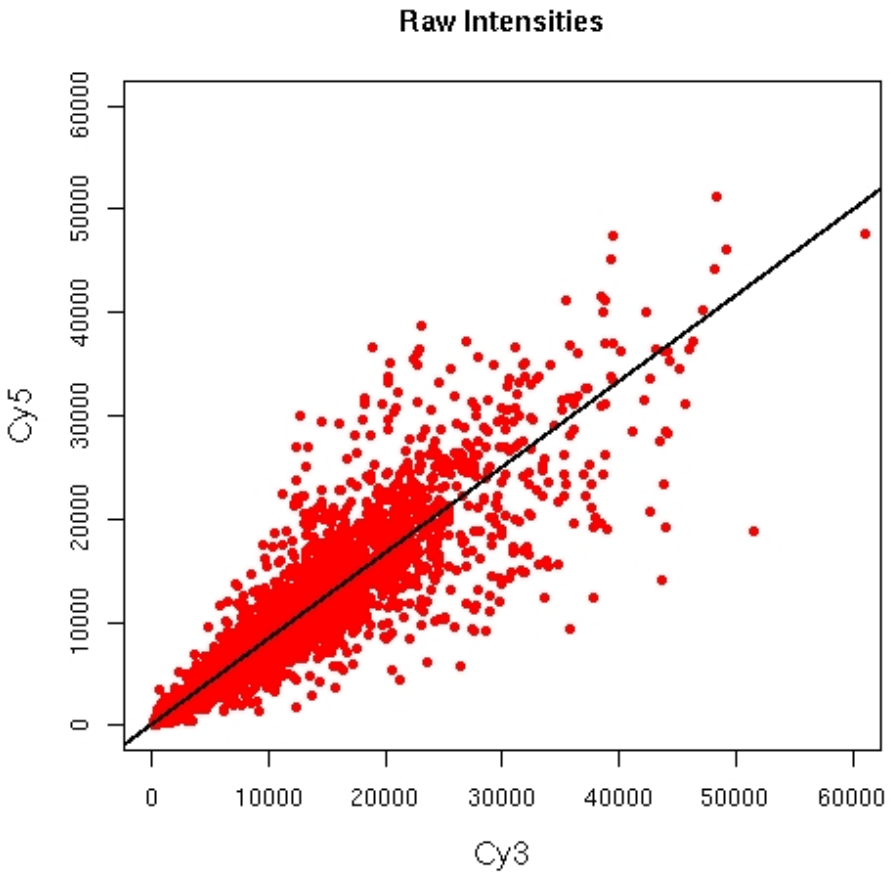
  Background subtraction can cause ratio inflation, therefore background corrected intensities below threshold are often set to threshold or similar value.

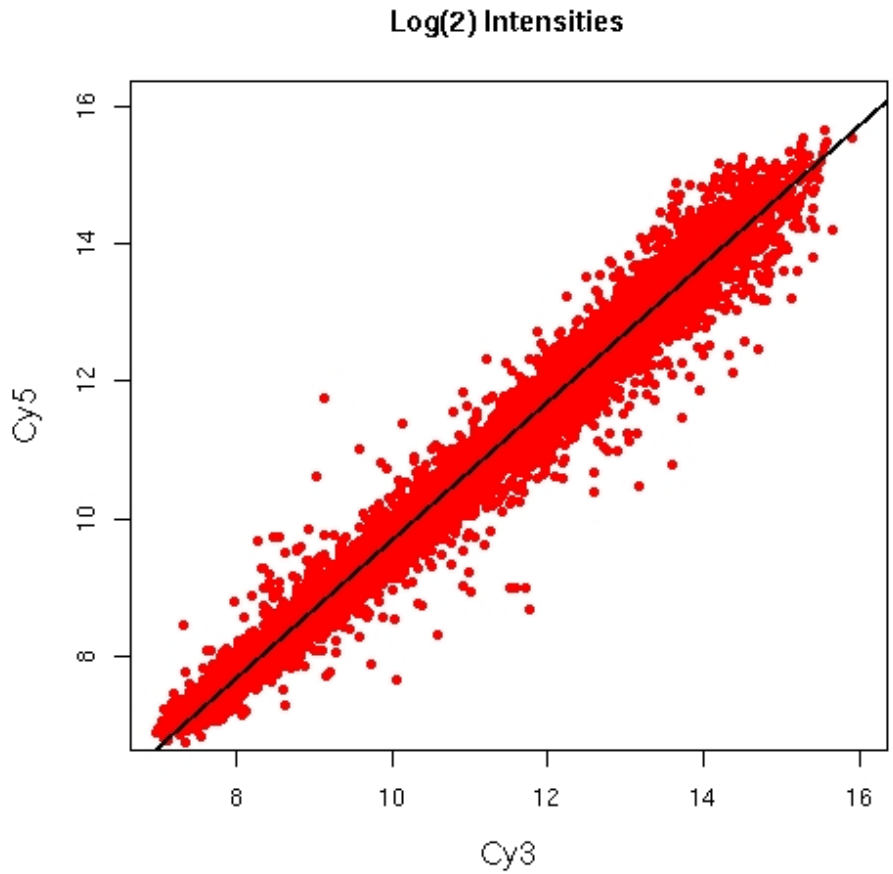
Normalization

Normalization is the process of balancing the intensities of the channels to account for variations in labeling and hybridization efficiencies. To achieve this, various adjustment strategies are used to force the distribution of all ratios to have a median (mean) of 1 or the log-ratios to have a median (mean) of 0.



Log Transformation: Scatter Plots





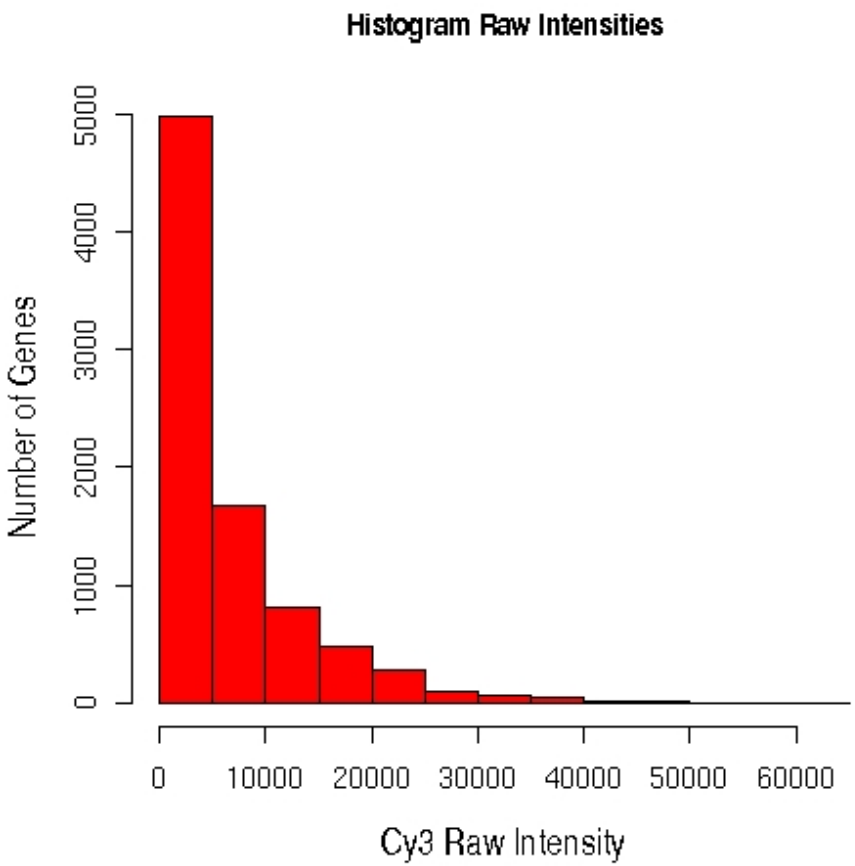
Reasons for working with log-transformed intensities and ratios

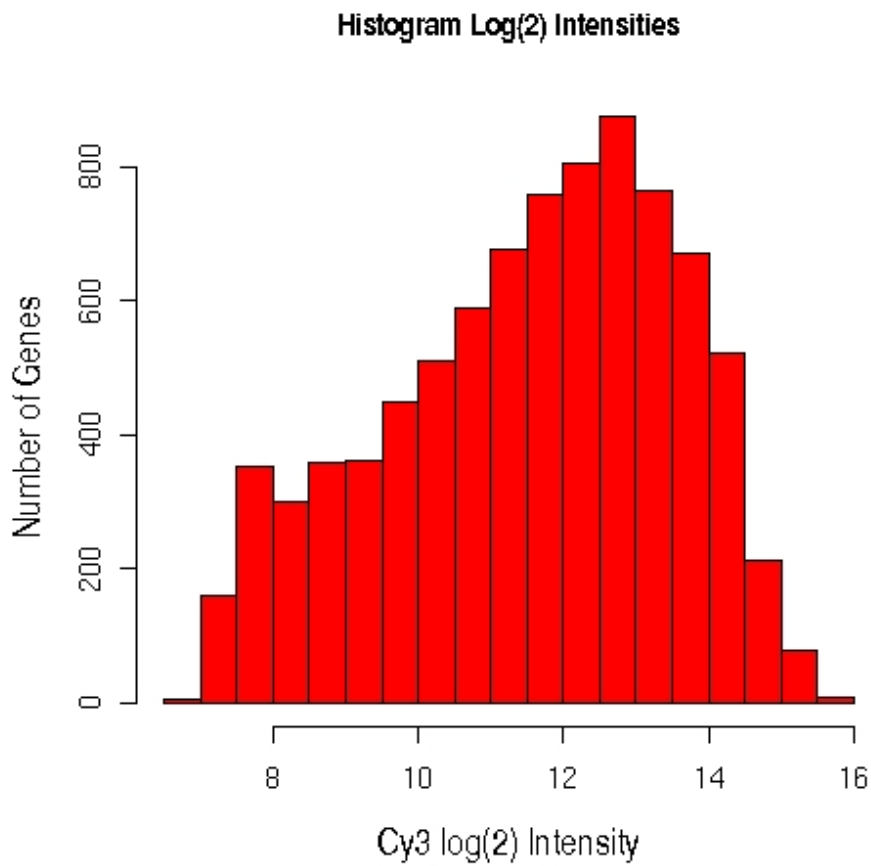
(1) spreads features more evenly across intensity range

(2) makes variability more constant across intensity range

(3) results in close to normal distribution of intensities and experimental errors

Log Transformation: Histograms





Distribution of log transformed data is closer to being bell-shaped

Normalization If Large Fraction of Genes IS DE

Minimize normalization requirements (dynamic range limits)

- Pre-scanning: hybridize equal amounts of label
- During scanning: balance average intensities through laser power and PMP adjustments

Normalization if large fraction of genes is DE

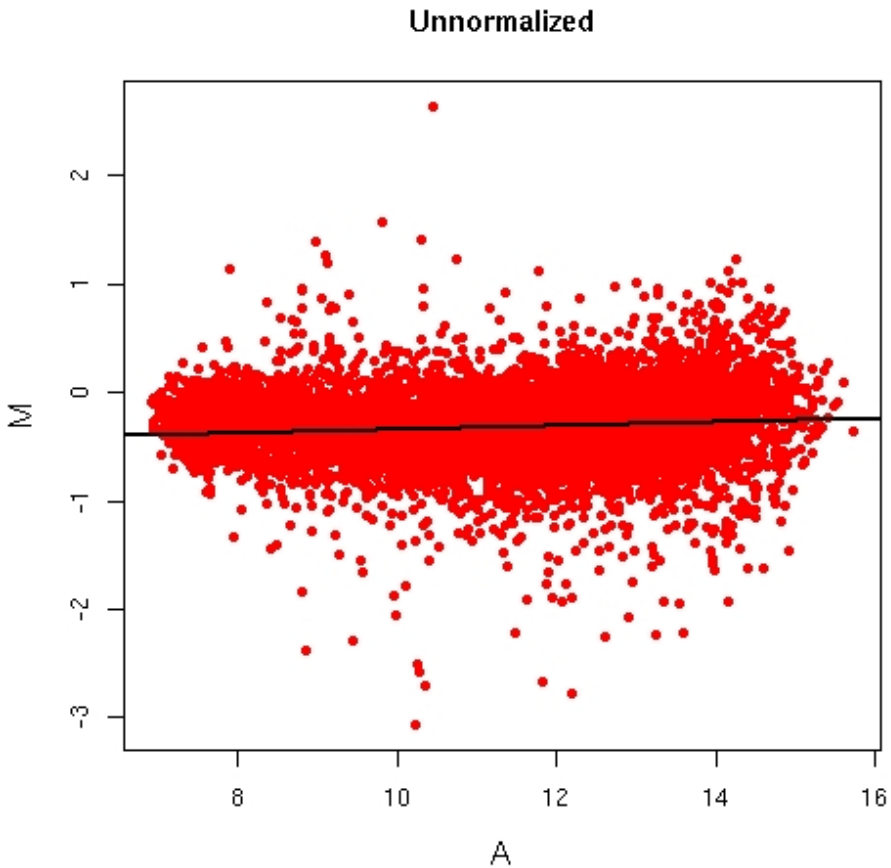
- Spike-in controls
- Housekeeping controls
- Determine constant feature set

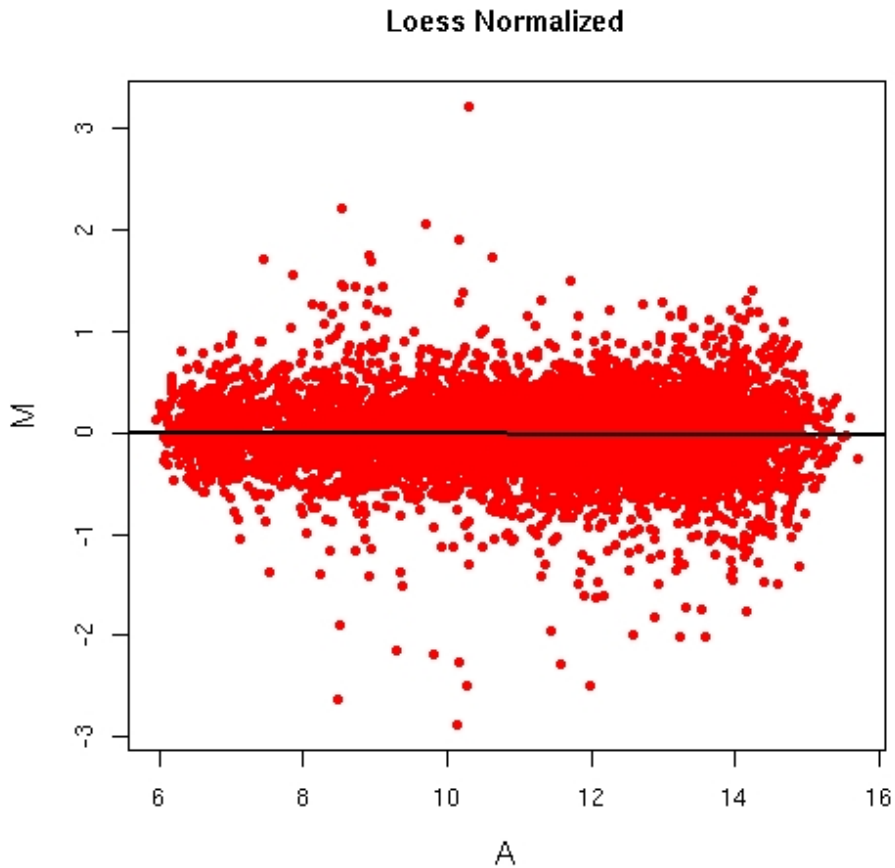
Normalization If Large Fraction of Genes IS NOT DE

Global Within-Array Normalization

- Multiply one channels with normalization factor
  $\Rightarrow$ Ch2 x mCh1/mCh2 (treats both channels differently)
- Linear regression fit of log2(Ch2) against log2(Ch1)
  $\Rightarrow$ adjust Ch1 with fitted values (treats both channels differently)
- Linear regression fit of log2(ratios) against avg log2(int)
  $\Rightarrow$ subtract fitted value from raw log ratios (treats both channels equally)
- Non-linear regression fit of log2(ratios) against avg log2(int)
  Most commonly used: Loess (locally weighted polynomial) regression joins local regressions with overlapping windows to smooth curve
  $\Rightarrow$ subtract fitted value on Loess regression from raw log ratios (treats both channels equally)

MA Plots


Unnormalized
M
A
-3
-2
-1
0
1
2
8
10
12
14
16


Loess Normalized
M
A
-3
-2
-1
0
1
2
3
6
8
10
12
14
16

Normalization If Large Fraction of Genes IS NOT DE

Spacial Within-Array Normalization

All of the above methods can be used to correct for spacial bias on the array. Examples:

- Block or Print Tip Loess
- 2D Loess Regression

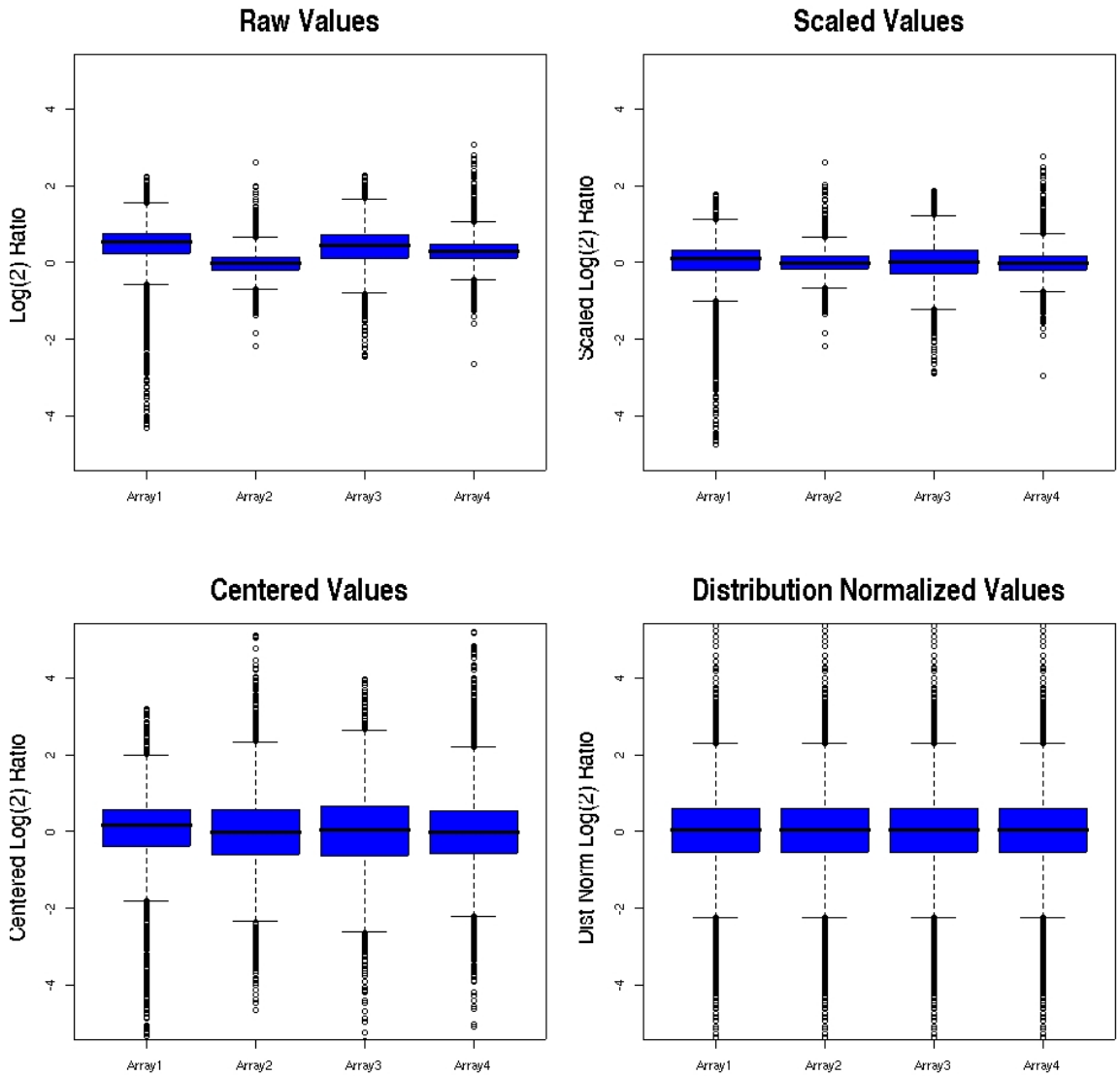
Normalization If Large Fraction of Genes IS NOT DE

Between-Array Normalization

To compare ratios between dual-color arrays or intensities between single-color arrays

- Scaling
  ⇒ log(rat) - mean log(rat) or log(int) - mean log(int)
  ⇒ Result: mean = 0
- Centering (z-value)
  ⇒ [rat - mean(rat)] / [STD] or [int - mean(int)] / [STD]
  ⇒ Result: mean = 0, STD = 1
- Distribution Normalization (apply to group of arrays!)
  ⇒ (1) Generate centered data, (2) sort each array by intensities, (3) calculate mean for sorted values across arrays, (4) replace sorted array intensities by corresponding mean values, (5) sort data back to original order
  ⇒ Result: mean = 0, STD = 1, identical distribution between arrays

Box Plots for Between-Array Normalization Steps



Analysis Methods for Affymetrix Gene Chips

Method	BG Adjust	Normalization	MM Correct	Probeset Summary
MAS5	regional adjustment	scaling by constant	subtract idealized MM	Tukey biweight average
gcRMA	by GC content	quantile normalization	/	robust fit of linear model
RMA	array background	quantile normalization	/	robust fit of linear model
VSN	/	variance stabilizing TF	/	robust fit of linear model
dChip	/	by invariant set	/	multiplicative model
dChip.mm	/	by invariant set	subtract mismatch	multiplicative model

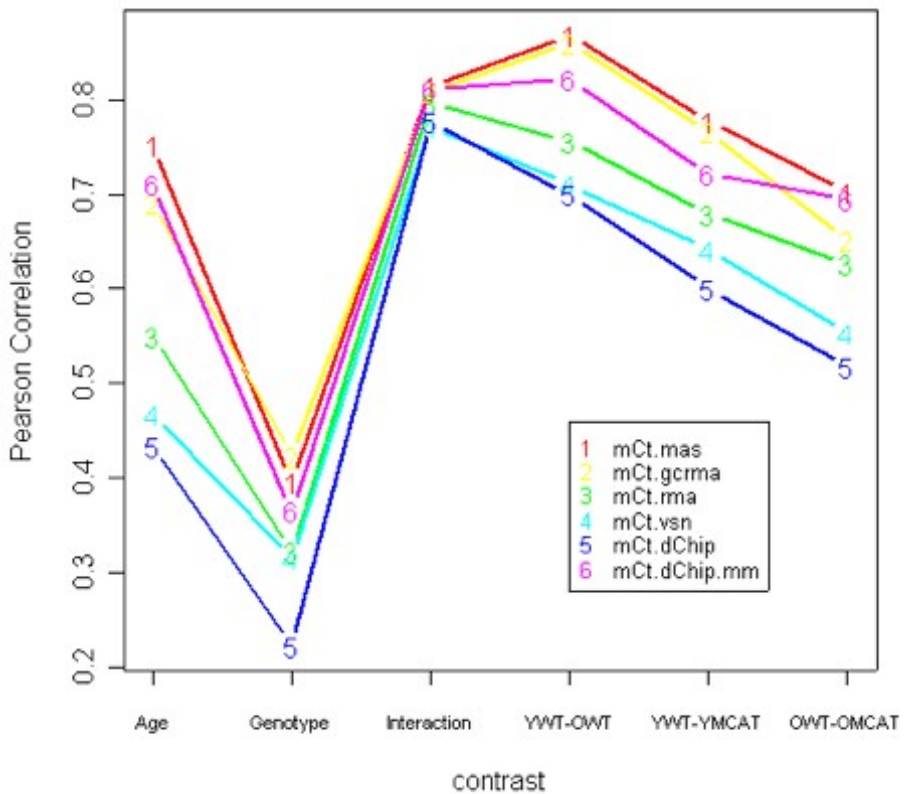
Qin et al. (2006), BMC Bioinfo, 7:23.

Reverences

MAS 5.0: Affymetrix Documentation: MAS5
PLIER: Affymetrix Documentation: PLIER, not included here
gcRMA: Wu et al. (2004), JASA, 99, 909-917.
RMA: Irizarry et al. (2003), Nuc Acids Res, 31, e15.
VSN: Huber et al. (2002), Bioinformatics, 18, Suppl I S96-104.
dChip & dChip.mm: Li & Wong (2001), PNAS, 98, 31-36.

Performance Comparison of Affy Methods

Qin et al. (2006), BMC Bioinfo, 7:23: 24 RNA samples hybridized to chips and 47 genes tested by qRT-PCR, plot shows PCC for 6 summary contrasts of 6 methods.



MAS5, gcRMA, and dChip (PM-MM) outperform the other methods. PLIER not included here.

Analysis of Differentially Expressed Genes

- Advantages of statistical test over fold change threshold for selecting DE genes
  - Incorporates variation between measurements
  - Estimate for error rate
  - Detection of minor changes
  - Ranking of DE genes
- Approaches
  - Parametric test: t-test
  - Non-parametric tests: Wilcoxon sign-rank/rank-sum tests
  - Bootstrap analysis (boot package)
  - Significance Analysis of Microarrays (SAM)
  - Linear Models of Microarrays (LIMMA)
  - Rank Product
  - ANOVA and MANOVA (R/maanova)
- Multiplicity of testing: p-value adjustments
  - Methods: fdr, bonferroni, etc.

Outline

Microarray Databases and Depositories

- NCBI GEO: http://www.ncbi.nlm.nih.gov/geo
- Microarray @ EBI: http://www.ebi.ac.uk/microarray
- SMD: http://genome-www5.stanford.edu
- Many Others

Outline

Why Using R and BioConductor for Array Analysis?

- Complete statistical package and programming language
- Useful for all bioscience areas
- Powerful graphics
- Access to fast growing number of analysis packages
- Is standard for data mining and biostatistical analysis
- Technical advantages: free, open-source, available for all OSs

Books & Documentation

- simpleR - Using R for Introductory Statistics (Gentleman et al., 2005)
- Bioinformatics and Computational Biology Solutions Using R and Bioconductor (John Verzani, 2004)
- UCR Manual (Thomas Girke)

Installation

1. Install R binary for your operating system from:

   http://cran.at.r-project.org

2. Install the required packages from BioConductor by executing the following commands in R:

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite()
> biocLite(c("GOstats", "Ruuid", "graph", "GO", "Category",
"plier", "affylmGUI", "limmaGUI", "simpleaffy",
"ath1121501", "ath1121501cdf", "ath1121501probe", "biomaRt",
"affycoretools"))
```

R Essentials

```
# General R command syntax
> object <- function(arguments)

# Execute an R script
> source("homework_script.R")

# Finding help
> ?function

# Load a library
> library(affy)

# Summary of all functions within a library
> library(help=affy)

# Load library manual (PDF file)
> openVignette()
```

Outline

Obtain Sample Data from GEO

- Retieve the *"Arabidopsis light treatment series"* (GSE5617) from GEO with the following query:

  `Arabidopsis[Organism] AND Atgenexpress[Title] AND light[Title]`

- Download the following Cel files from this GSE5617 series:

  `GSM131177.CEL`
  `GSM131192.CEL`
  `GSM131207.CEL`
  `GSM131179.CEL`
  `GSM131193.CEL`
  `GSM131209.CEL`
  `GSM131181.CEL`
  `GSM131195.CEL`
  `GSM131211.CEL`

  Batch download:

  GEO_CEL.zip

Define Replicates and Treatments

- Generate targets.txt file and save it in your working directory. It should contain the following content:

Name	FileName	Target
DS_REP1	GSM131177.CEL	dark45m
DS_REP2	GSM131192.CEL	dark45m
DS_REP3	GSM131207.CEL	dark45m
PS_REP1	GSM131179.CEL	red1m_dark44m
PS_REP2	GSM131193.CEL	red1m_dark44m
PS_REP3	GSM131209.CEL	red1m_dark44m
BS_REP1	GSM131181.CEL	blue45m
BS_REP2	GSM131195.CEL	blue45m
BS_REP3	GSM131211.CEL	blue45m

Homework Tasks

A. Generate expression data with RMA, GCRMA and MAS 5.0. Create box plots for the raw data and the RMA normalized data.

B. Perform the DEG analysis with the limma package and determine the differentially expressed genes for each normalization data set using as cutoff an adjusted p-value of $\leq$0.05. Record the number of DEGs for each of the three normalization methods in a summary table.

C. Create for the DEG sets of the three sample comparisons a venn diagram (adjusted p-value cutoff $\leq$0.05).

D. Generate a list of genes (probe sets) that appear in all three filtered DEG sets (from B.).

$\Rightarrow$ Command summary: source("homework_script.R")

R Commands for Normalization

```
# Load required libraries
> library(affy); library(limma); library(gcrma)

# Open limma manual
> limmaUsersGuide()

# Import experiment design information from targets.txt
> targets <- readTargets("targets.txt")

# Import expression raw data and store them in AffyBatch object
> data <- ReadAffy(filenames=targets$FileName)

# Normalize the data with the RMA method and store results in exprSet
object
> eset <- rma(data) # RMA and GCRMA store log2 intensities and MAS5
absolute intensities.

# Print the analyzed file names
> pData(eset)

# Export all affy expression values to a tab delimited text file
> write.exprs(eset, file="affy_all.xls")
```

R Commands for Differential Expression Analysis

```
# Create appropriate design matrix and assign column names
> design <- model.matrix(~ -1+factor(c(1,1,1,2,2,2,3,3,3)));
colnames(design) <- c("S1", "S2", "S3")

# Create appropriate contrast matrix for pairwise comparisons
> contrast.matrix <- makeContrasts(S2-S1, S3-S2, S3-S1, levels=design)

# Fit a linear model for each gene based on the given series of arrays
> fit <- lmFit(eset, design)

# Compute estimated coefficients and standard errors for a given set of
contrasts
> fit2 <- contrasts.fit(fit, contrast.matrix)

# Compute moderated t-statistics and log-odds of differential expression
by empirical Bayes shrinkage of the standard errors towards a common
value
> fit2 <- eBayes(fit2)

# Generate list of top 10 DEGs for first comparison
> topTable(fit2, coef=1, adjust="fdr", sort.by="B", number=10)
```

Online Manual

Continue on online manual.