Custom microarray analysis in R

Matt Settles

University of Idaho
Bioinformatics and Computational Biology Program

February 4, 2013

Outline

- Experiment_Directory
  - Raw_Data
  - Design_Files
  - Figures
  - Tables
  - Data
  - Analysis.R
  - Venn.R
  - functions.R
  - targets.txt

targets.txt - an annotated data frame

An annotated data frame allows you to manage the samples for the experiment and includes at minimum sample names with associated raw data files. The framework however is flexible and allows for more information about samples to be included.

See targets.txt file in example data

Components of Raw Data

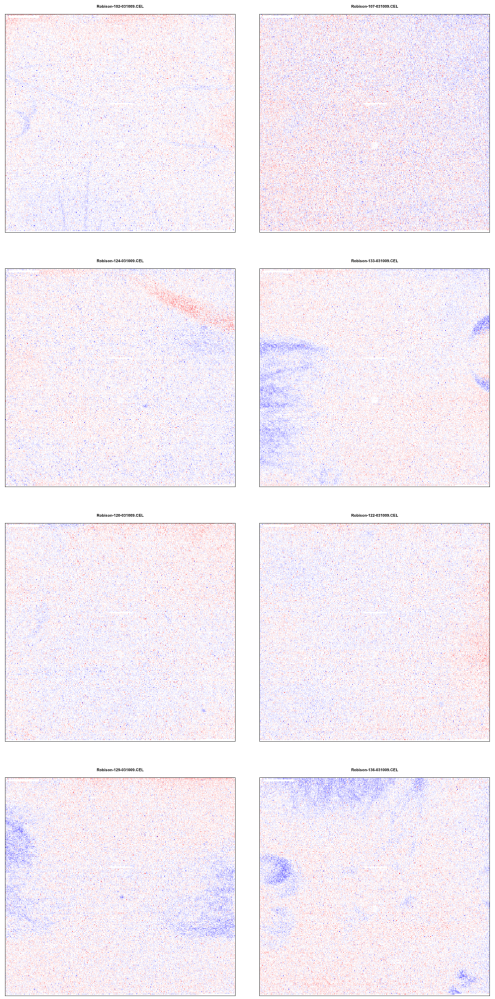
- Quantified intensities for each microarray probe (spot), usually identified by X,Y coordinate
  - Affymetrix: binary CEL files
  - Nimblegen: plain text pair or xys files
  - Agilent: plain text gpr files
- Microarray description file, associates probes to probesets and genes
  - Affymetrix: CEL description files or CDF files (Bioconductor provides)
  - Nimblegen: Nimblegen description files or NDF files (must create your own packages)
  - Agilent: GAL files (can be used in Limma)

Quality Assurance

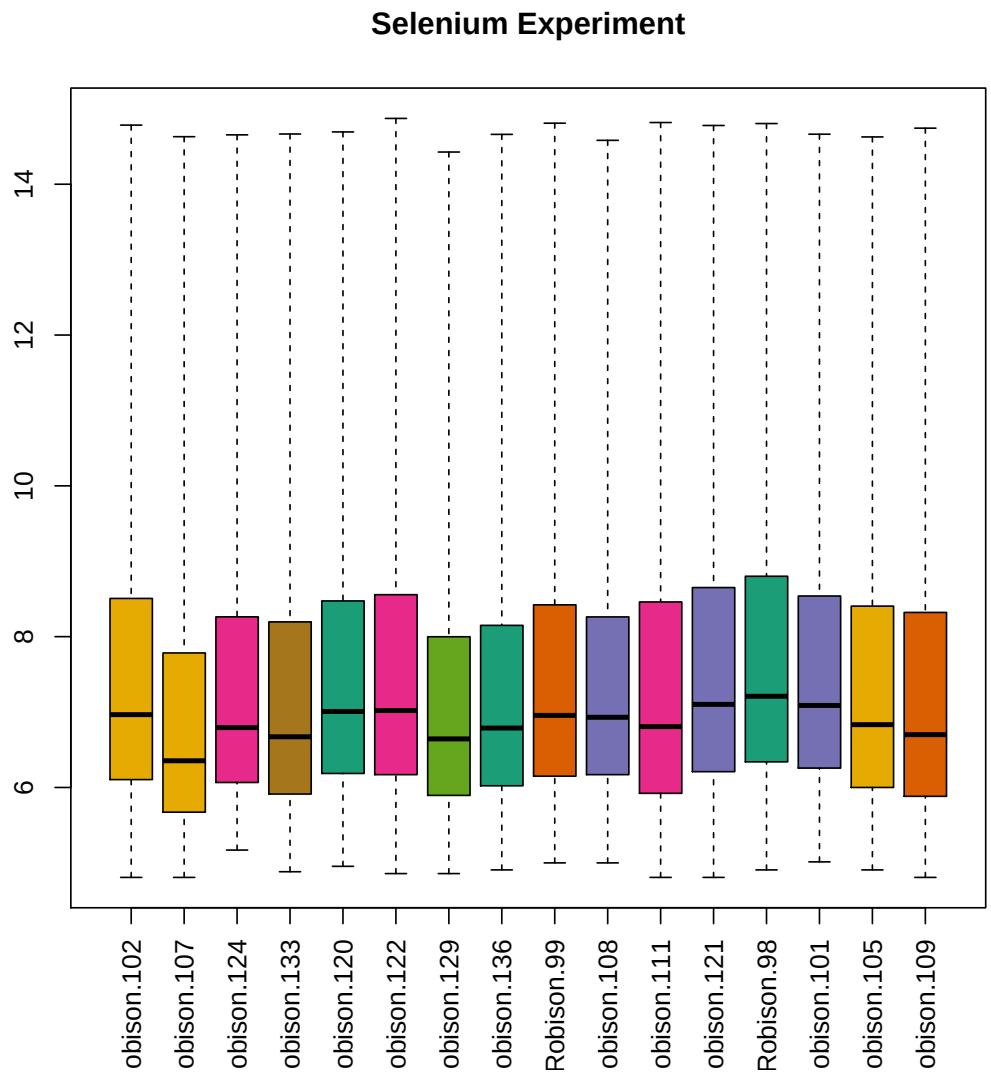
1. Visual check of chip images or pseudo-images.
2. Consistency across microarrays.
3. Observed patterns meet expectations.

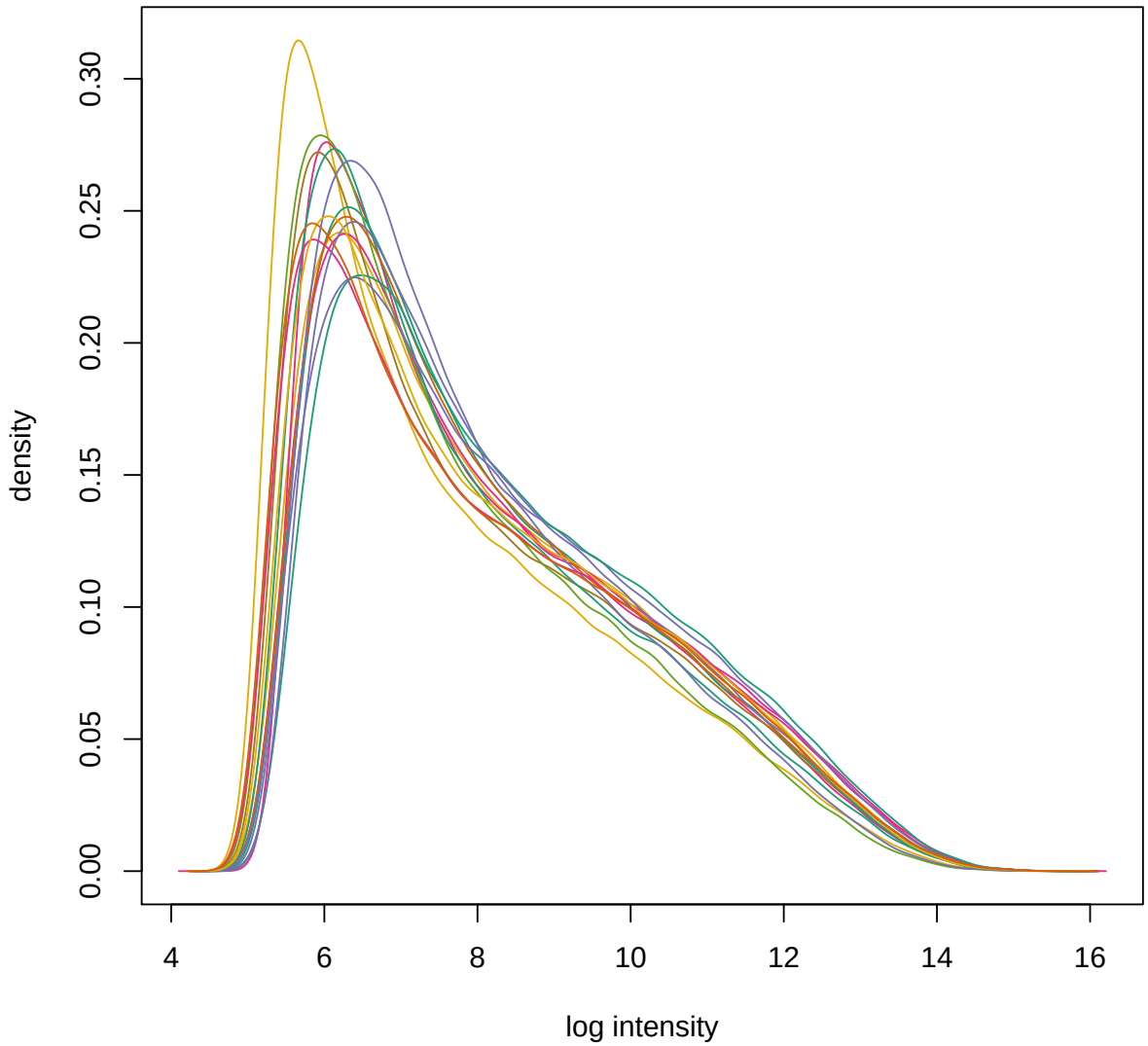
Pseudo-chip images

- Pseudo-chip images are created from the data themselves and not from the original scanned images
- Can be created from the raw intensities, log intensities, model fitting residuals and other
- Looking for large anomalies

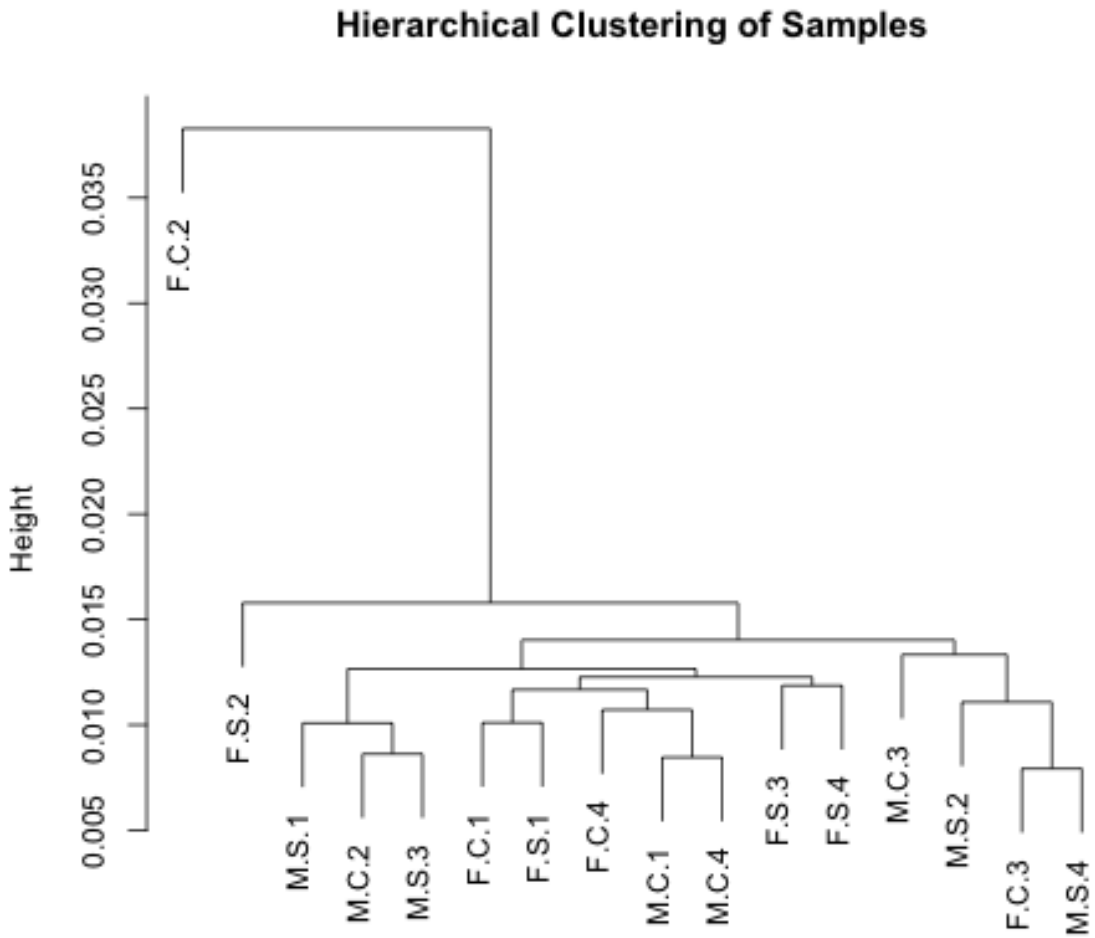


Consistency across microarrays




density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
4
6
8
10
12
14
16
log intensity

Expectations



Preprocessing

The goals of preprocessing Affymetrix microarray data are three fold:

1. To remove variation due to technical sources, while preserving variation from biological sources.
2. To normalize a set of microarrays in order to make them comparable.
3. To produce summarized expression values for each probe set.

Background correction Local Effects. Deviations from actual expression levels are introduced by many sources including scanner artifacts, non-specific binding (hybridization), RNA quality, reagents, etc.. All of which can be considered as background noise.

Probe level normalization Array Effects. The role of normalization is to compensate for the technical effects, while preserving the effects due to the biology.

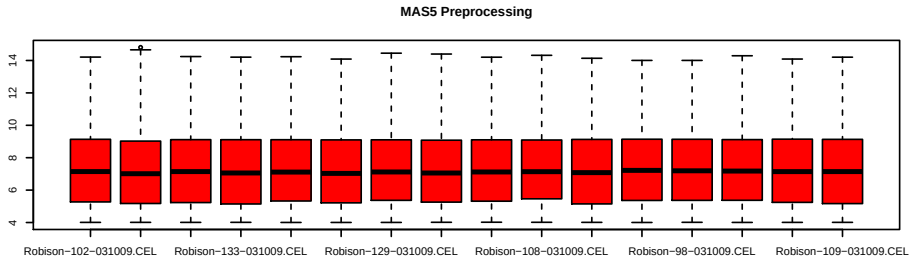
PM Correction PM correction routines are designed to account for non-specific binding by use of the MM probes.

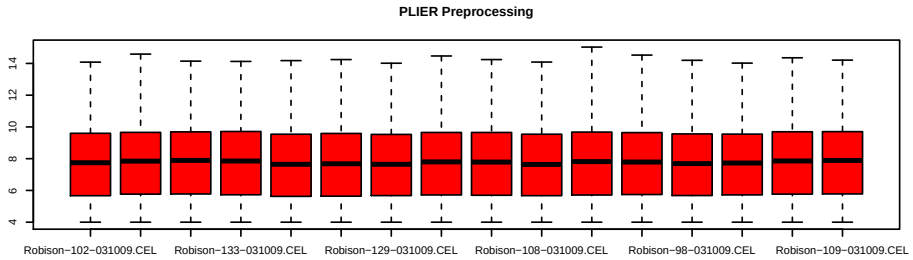
Probe set summarization Probe set summarization produces a single value for a probe set that is the estimated expression level for the probe set (i.e. gene).

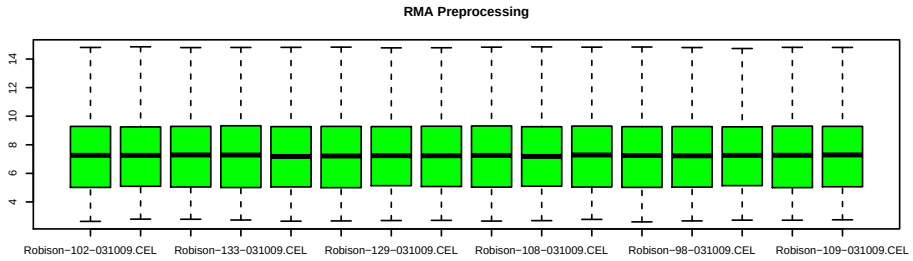
Probe set normalization Same purpose as Probe level normalization but on the probe sets.

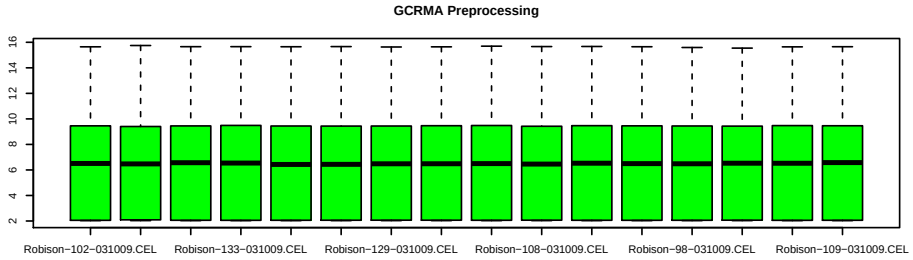
Table : Breakdown of the preprocessing steps for the MAS5, Plier, RMA, GCRMA, and dChip preprocessing pipelines.

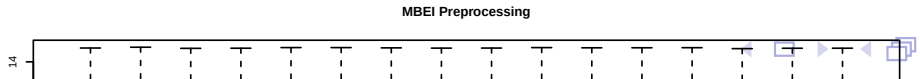
	MAS5	**Plier**	**RMA**	**GCRMA**	**dChip**
Background Correction	weighted average	none	RMA (global model)	GCRMA (model based)	none
Probe Level Normalization	none	quantile normalization	quantile normalization	quantile normalization	invariant set
PM Correction	ideal mismatch	none	none	none	subtract MM none
Probe set Summarization	Tukey biweight	Plier	median polish	median polish	MBEI
Probe set Normalization	mean scaled	none	none	none	none

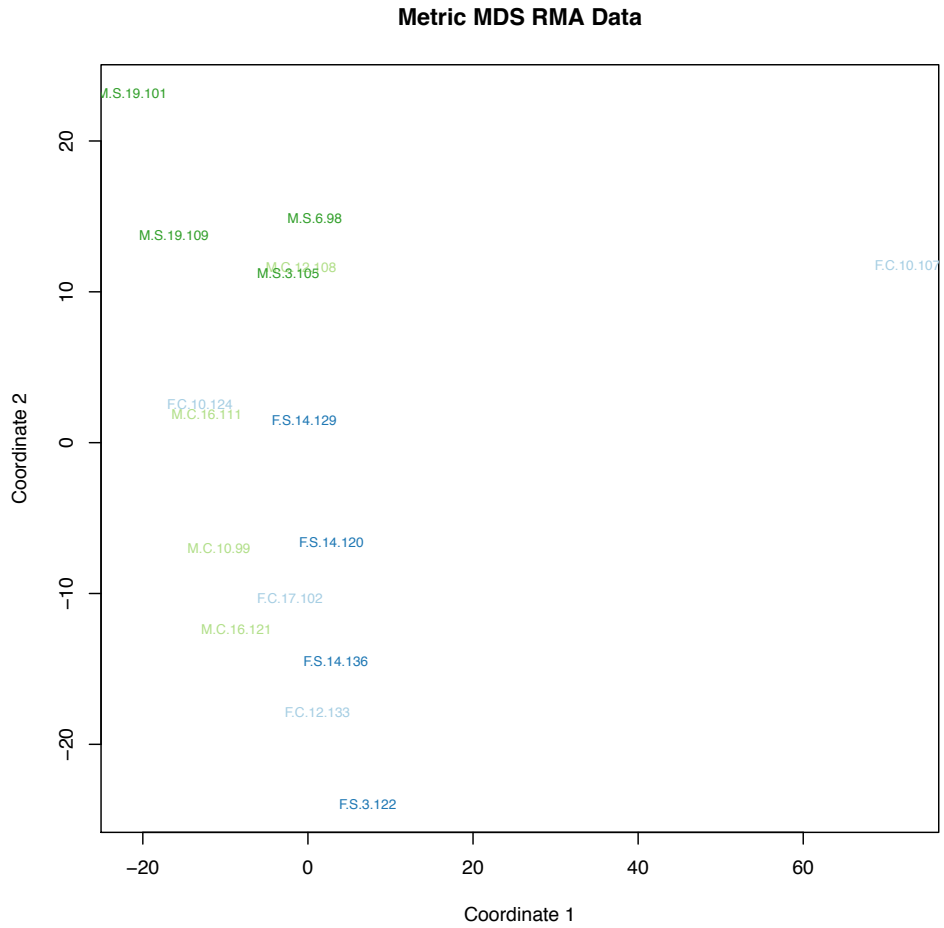

MAS5 Preprocessing
Robison−102−031009.CEL
Robison−133−031009.CEL
Robison−129−031009.CEL
Robison−108−031009.CEL
Robison−98−031009.CEL
Robison−109−031009.CEL


PLIER Preprocessing
Robison−102−031009.CEL
Robison−133−031009.CEL
Robison−129−031009.CEL
Robison−108−031009.CEL
Robison−98−031009.CEL
Robison−109−031009.CEL


RMA Preprocessing
Robison−102−031009.CEL
Robison−133−031009.CEL
Robison−129−031009.CEL
Robison−108−031009.CEL
Robison−98−031009.CEL
Robison−109−031009.CEL


GCRMA Preprocessing
Robison−102−031009.CEL
Robison−133−031009.CEL
Robison−129−031009.CEL
Robison−108−031009.CEL
Robison−98−031009.CEL
Robison−109−031009.CEL


MBEI Preprocessing


Metric MDS RMA Data
M.S.19.101
M.S.6.98
M.S.19.109
M.S.3.105
F.C.10.107
F.C.10.124
M.C.16.111
F.S.14.129
M.C.10.99
F.S.14.120
F.C.17.102
M.C.16.121
F.S.14.136
F.C.12.133
F.S.3.122
Coordinate 2
20
10
0
−10
−20
−20
0
20
40
60
Coordinate 1

RMA

- Irizarry et al. (2003) Biostatistics 4(2):249-264.
- Irizarry et al. (2003) Nucleic Acids Research 31(4):e15.
- Bolstad et al. (2003) Bioinformatics 19(2):185-193.

RMA-Convolutation Background Correction

$$PM_{ijk} = \underbrace{bg_{ijk}}_{\uparrow} + \underbrace{s_{ijk}}_{\text{Signal for probe j of probe set k on array i}}$$

Background caused by optical noise and non-specific binding

$$B(PM_{ijk}) = E[s_{ijk} \mid PM_{ijk}] > 0$$

$$s_{ijk} \sim Exp(\lambda_{ijk}) \qquad bg_{ijk} \sim N(\beta_i, \sigma_i^2)$$

} Gives a closed-form transformation B()

RMA-Quantile Normalization

A technique for making two or more distributions identical in statistical properties.

- Each vector to be normalized must be the same length
- Sort the vectors from smallest to greatest.
- Compute the means at each sorted column, this is the new distribution.
- Unsort the original vectors using the new computed distribution.

Note: Assumes, they should have the same underlying distributions.

RMA-Median Polosh

Robust method suggested by Tukey for estimating the mean, row and column parameters of the model

$$Y_{ijk} = \underbrace{\mu_{ik}}_{\text{Log-scale expression level for gene k on array i}} + \underbrace{\alpha_{jk}}_{\text{Probe affinity effect; for each k, } \sum_j \alpha_{jk} = 0} + \varepsilon_{ijk}$$

Gene filtering

Purpose is to remove probe sets (aka genes) that are likely to be of no use (uninteresting) when modelling the data.

non-specific Choose genes to remove (or keep) without using any phenotypic variables in the filtering process. Result can then be used with any downstream process without bias.

specific Use of experimental variables to choose which genes should be kept/removed. Results can be used in a descriptive nature, should not be used for statistical testing.

Non-specific filters I

Annotation Based Filtering Can reduce be a predetermined set of genes (aka entrez ids), by Gene Ontology group, filter based on available annotation data.

Duplicate Probe Removal Probes determined by annotation to be pointing to the same gene are compared, and only the probe with the highest variance value will be retained.

Variance Based Filtering Perform numerical cutoff-based filtering (aka IQR). The intention is to remove uninformative probe sets, representing genes that were not expressed at all or no change is occurring. Observations have shown that unexpressed genes are detected most reliably through their low variability across samples. IQR is robust to outliers, a default cutoff value of 0.5 is motivated by the rule of thumb that in many tissues only 40% of genes are expressed.

Differential Expression Analysis I

LIMMA Linear Models for Microarray Data approach. Limma applies a linear model to the data and uses model contrasts to extract specific comparisons of interest. It then uses multiple testing correction based techniques (FDR, FWER, etc.) for MT corrected p-values. Can also choose a fold change parameter, to ensure that called genes change at least a pre-specified amount.

Variance Stabilization

- When doing statistical test, we estimate the variance for each gene individually. This is fine, *if* we have enough replicates, but with few replicages (say 3-5 per group), these variances are highly variable.
- In a moderated statistic, the estimated gene-specific variance s_g^2 is replaced by a weighted average of s_g^2 and s_0^2, which is a global variance estimator obtained from pooling all genes.
- This produces an interpolation between the t-test and a fold-change criterion.

LIMMA performs an empirical Bayes estimate for s_0^2.

FWER vs FDR

Aim: For a given type I error rate α, use a procedure to select a set of "significant" genes that guarantees a type I error rate less than or equal to α.

FWER Familywise Error Rate - is the probability of making one or more false discoveries, or type I errors among all the hypotheses when performing multiple hypotheses tests. Stringent test such as Bonferonni.

FDR False Discovery Rate - In a list of statistically significant findings, FDR procedures are designed to control the expected proportion of incorrectly rejected null hypotheses ("false discoveries"). Less stringent test such as BH and q-value.

Multiple Testing Correction

Many methods available in the bioconductor packages *multtest* and *qvalue*.

Bonferonni Basically multiple raw p-value by the number of tests. **Unrealistic** given the number of genes tested.

Benjimini and Hochberg (BH) Controls the false discovery rate. Works for independent test statistics and for some types of dependence. Tends to be conservative if many genes are differentially expressed.

q-value the q-value of a gene is defined as the minimal estimated FDR at which it appears significant.

Public Repositories for Microarrays Data

GEO http://www.ncbi.nlm.nih.gov/geo/

ArrayExpress http://www.ebi.ac.uk/arrayexpress/