Making Sense of Complicated Microarray Data

Data normalization and transformation

Prashanth Vishwanath
Boston University

Dealing with Data

- Before any pattern analysis can be done, one must first normalize and filter the data.
- Normalization facilitates comparisons between arrays.
- Filtering transformations can eliminate questionable data and reduce complexity.

Why Normalize Data?

- Goal : Measure ratios of gene expression levels.
  - Ratio = T_i/C_i. Ratio of measured treatment intensity to control intensity for the i^{th} spot
  - In self/self experiments, ratio = 1.

But this is never true

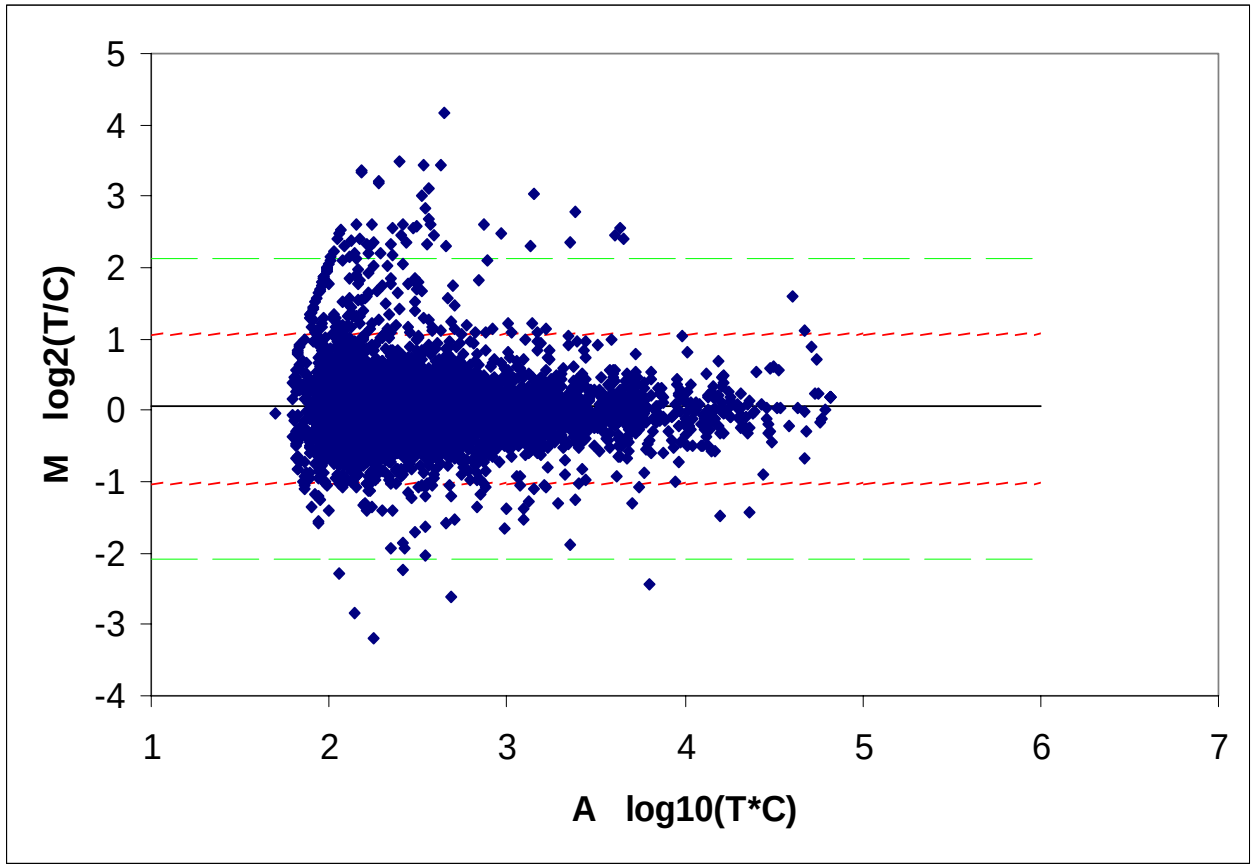
- Imbalances can be caused by
  - Different incorporation of dyes
  - Different amounts of mRNA
  - Different scanning parameters etc.
- Normalization balances red and green intensities.
  - Reduce intensity-dependent effects
  - Remove non-linear effects

The starting point - Using the $\log_2$ ratio

- The $\log_2$ ratio treats up and down regulated genes equally.
  - e.g. when looking for genes with more than 2 fold variation in expression

Differential expression	ratio is	log ratio is
over-expression	≥ 2	≥ 1
neutral	≈ 1	≈ 0
under-expression	≤ 0.5	≤ -1

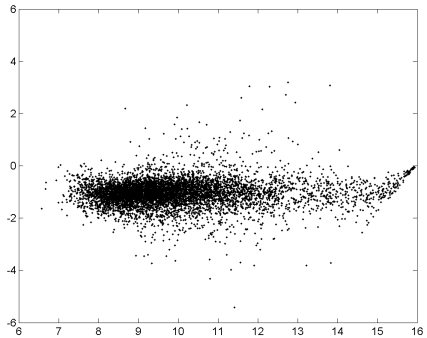
The MA or Ratio-Intensity plots



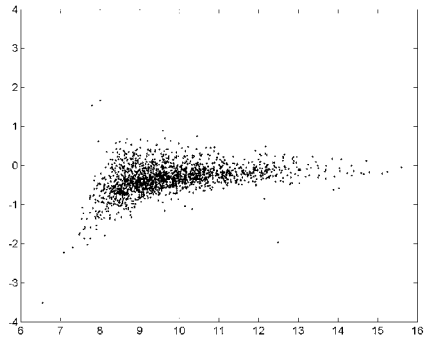
- Differential expression log ratio **M = $\log_2$(R / G)**
- Log intensity **A = $\log_{10}$(R*G) / 2**

Common problems diagnosed using RI plots

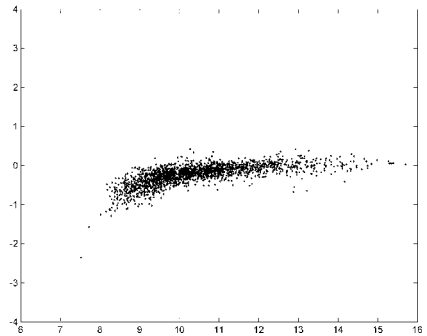
Saturation



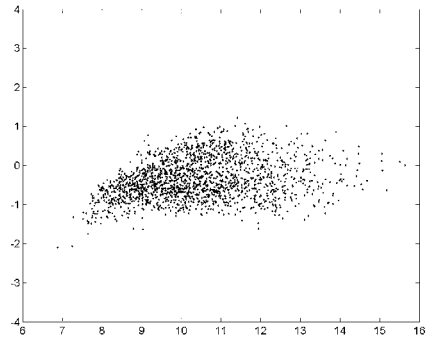
Low end variation



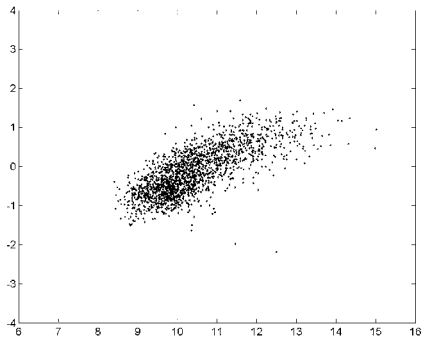
Curvature at Low intensity



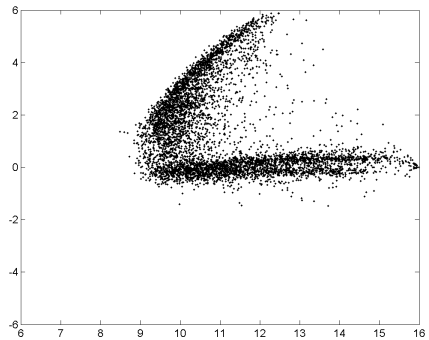
High end variation



Large curvature



Heterogeneity



Different Normalization techniques

- Total Intensity
- Iterative log(mean) centering
- Linear Regression
- Lowess correction

……..and others

- Dataset for normalization
  - Entire data set (all genes)
  - User defined dataset/ controls (e.g. housekeeping genes)

Total Intensity Normalization: mean or median

- Assumption: equal quantities of RNA in samples that are being compared

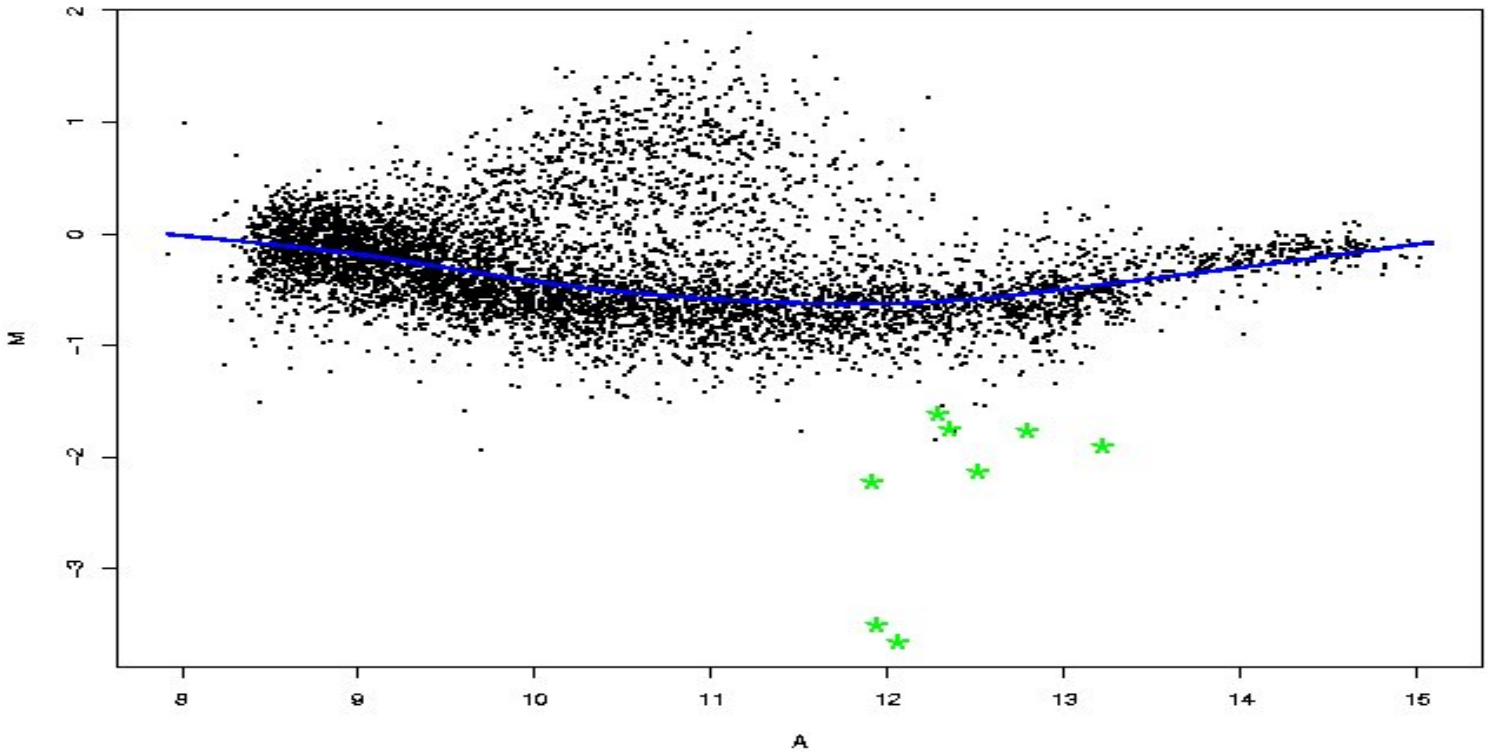
$$N_{total} = \frac{\sum_{g=1}^{Narray} Rg}{\sum_{g=1}^{Narray} Gg}$$

•Normalized expression ratio adjusts each ratio such that mean ratio is 1

$$\boldsymbol{log_2(ratio'_i) = log_2(ratio_i) - log_2(N_{total})}$$

A similar scaling factor could be used to normalize intensities across all arrays

Normalization - Lowess



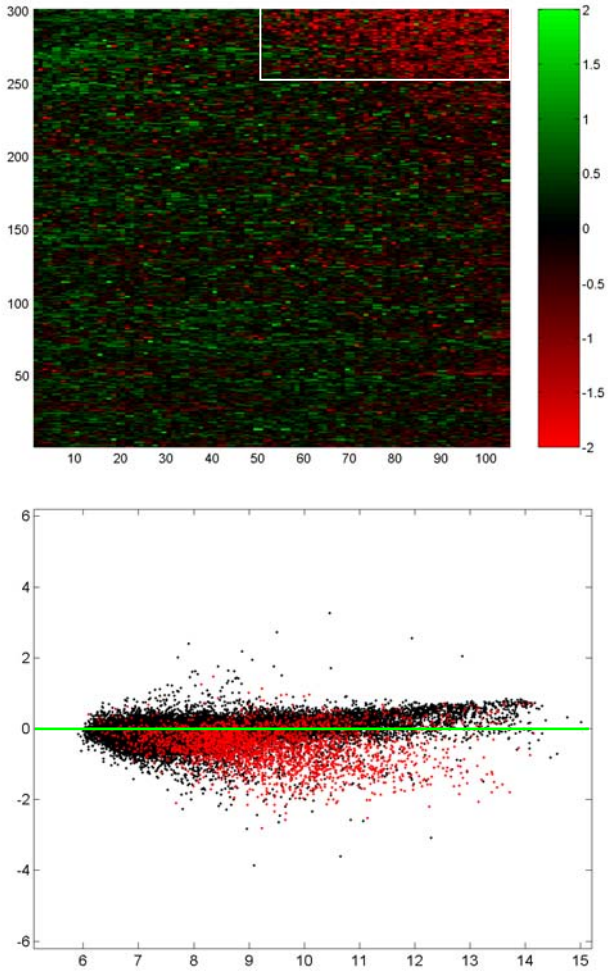
- Lowess detects systematic bias in the data
- Intensity-dependent structure
- Different ways to fit a lowess curve
  - –Local linear regression
  - –Splines

Global vs. local normalization

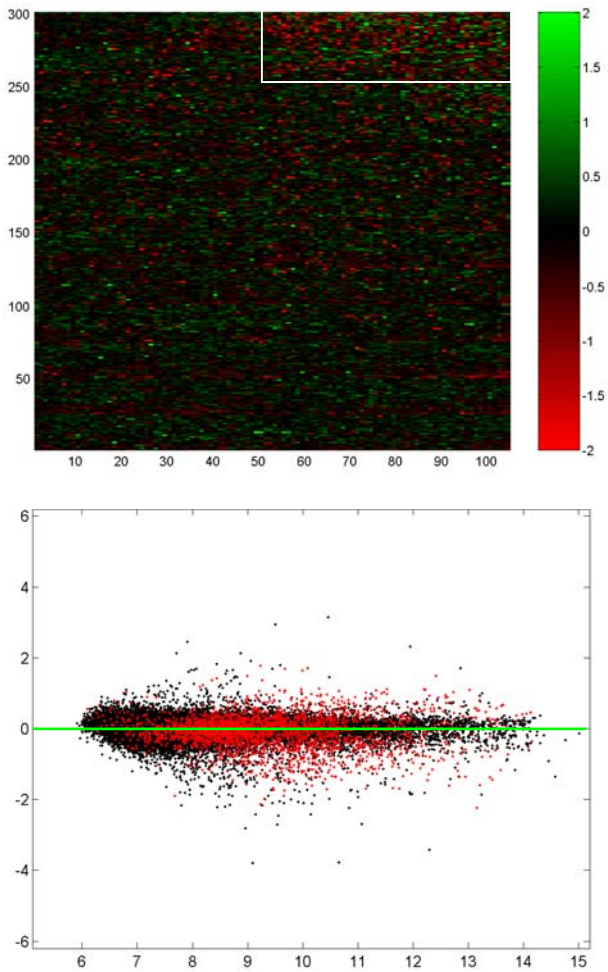
- Most normalization algorithms can be applied globally or locally
- Advantages of local normalization
  - Correct systematic spatial variation
    - Inconsistencies in the spotting pen
    - Variability in slide surface
    - Local differences in hybridization conditions
- An example of a local set may be each group of array elements spotted by a single spotting pen.

Spatial Lowess

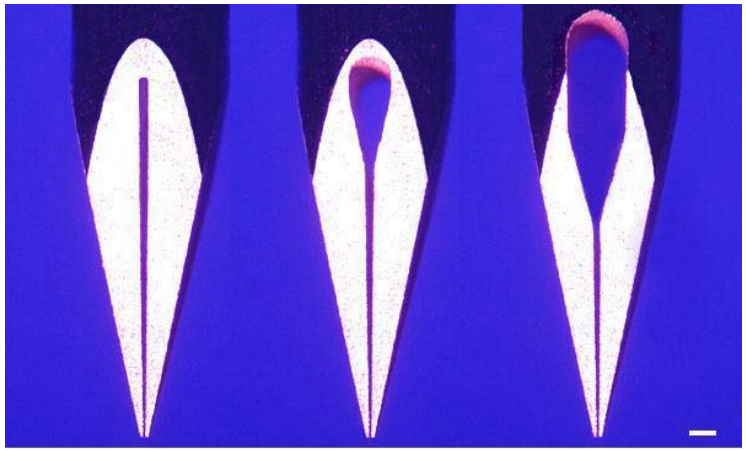
Before

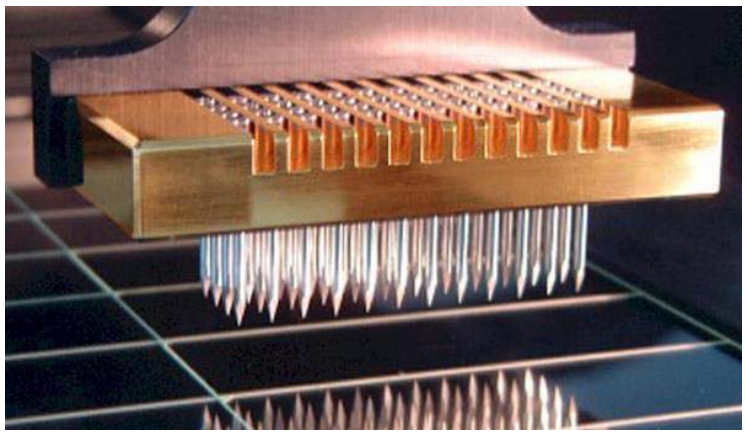


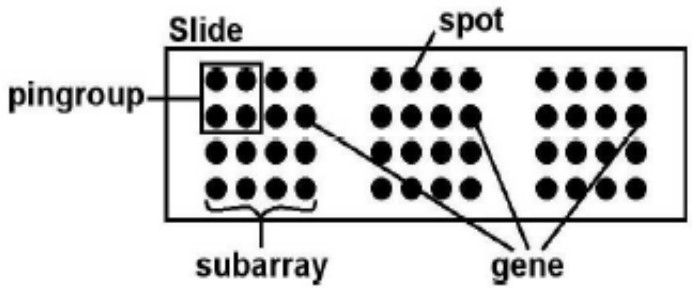
After spatial Lowess



Printing Variability

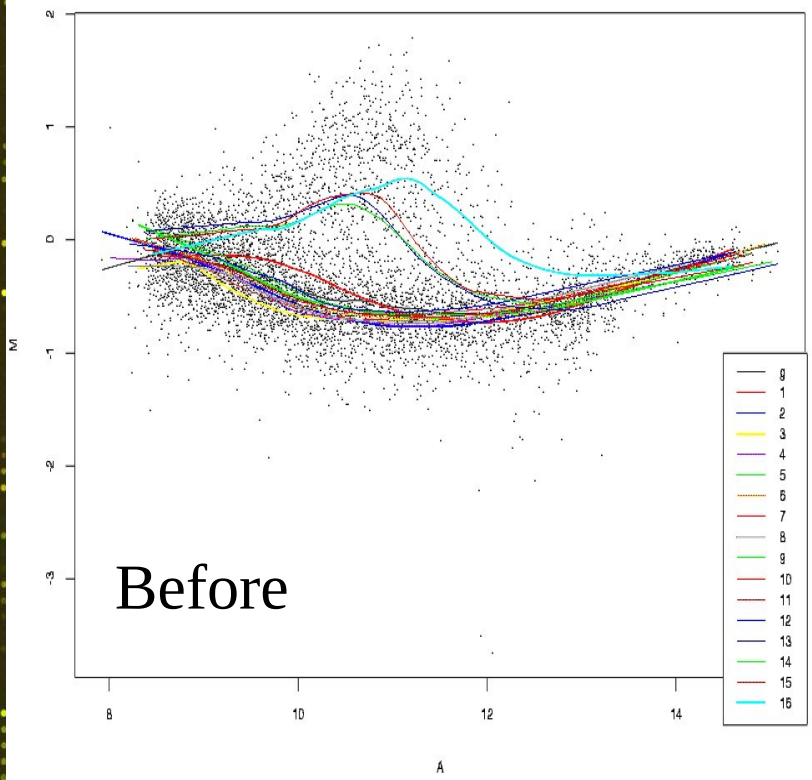


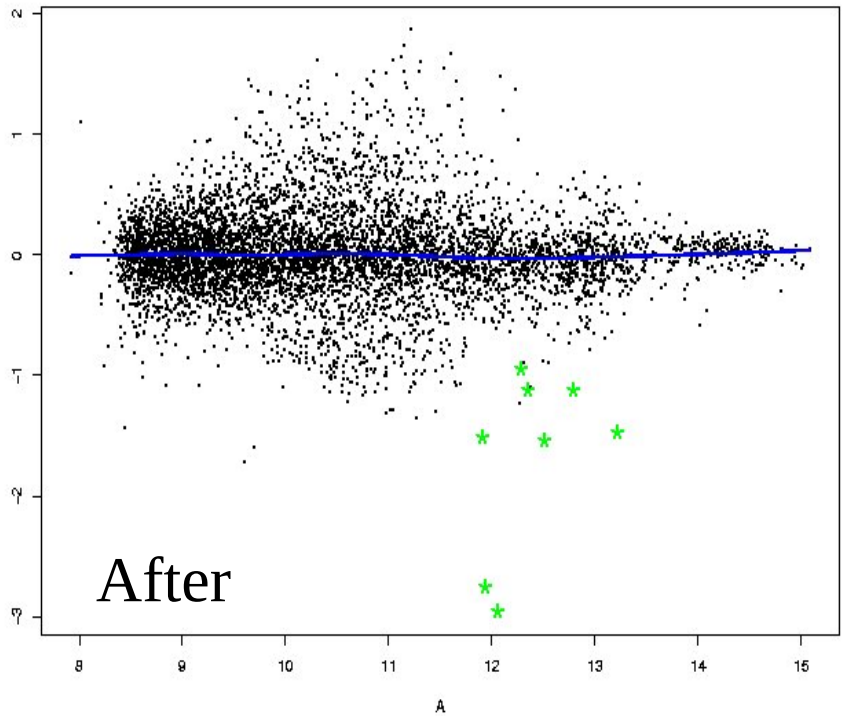




Normalization - print-tip-group

Assumption: For every print group, changes roughly symmetric at all intensities.





Variance regularization

- Stochastic processes can cause the variance of the measured $\log_2$(ratio) values to differ
- Adjust ratios such that variance is the same (using a single factor a_i for a subarray in a chip to scale all values in that subarray)
- Scaling factor = $\dfrac{\text{variance}_{\text{subarray}}}{\text{variance of all subarrays}}$

A similar measure can be used to regularize variances between arrays

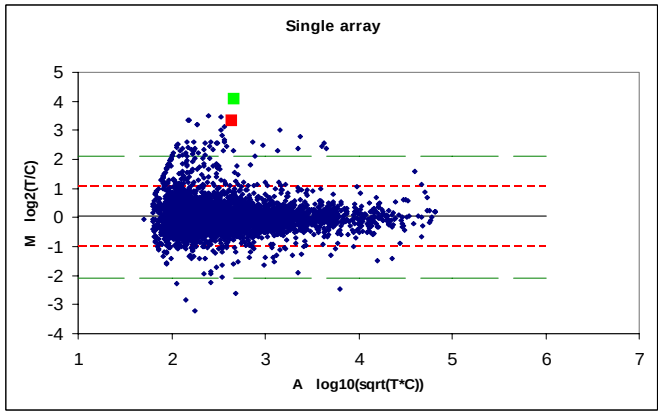
Averaging over replicates

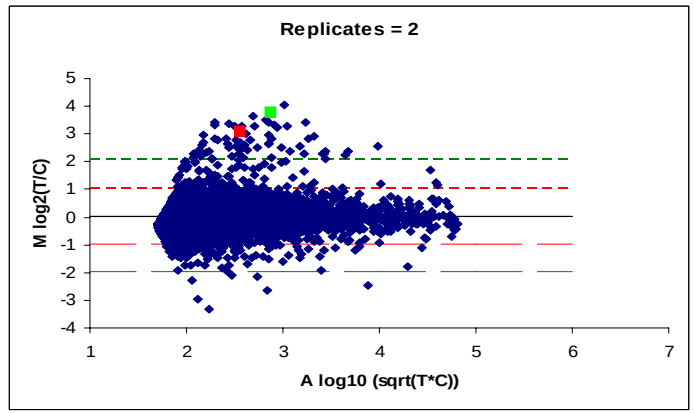
- Result is equivalent to taking the geometric mean

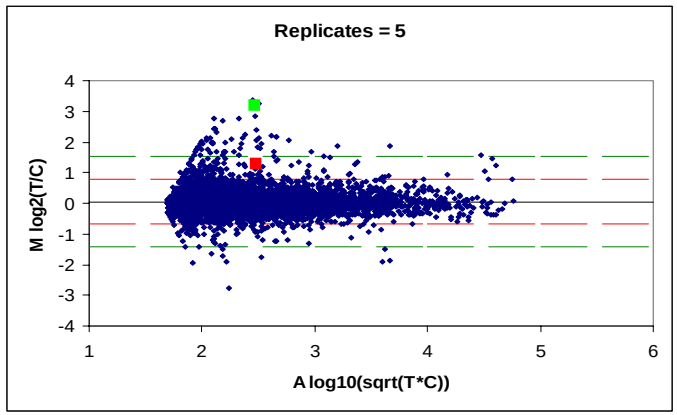
$$\text{ratio} = \frac{(T_{i1} * T_{i2} * T_{i3})^{1/3}}{(C_{i1} * C_{i2} * C_{i3})^{1/3}}$$

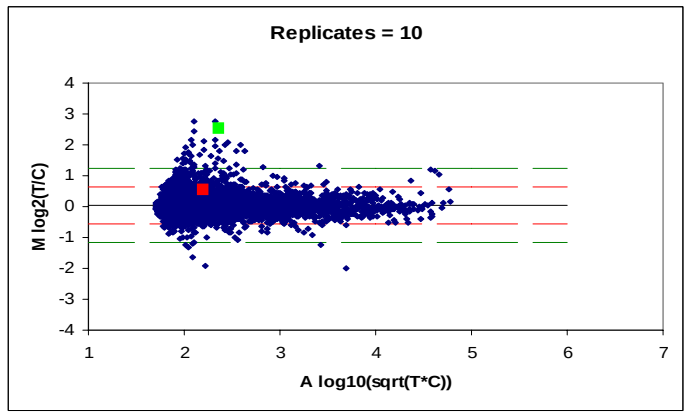
- Combining intensity data
  - Ratio of the geometric means of the channel intensities
- Dye swaps – helps in reducing dye effects

Number of replicates









- Integrin alpha 2b
- Pro-platelet basic protein

I have normalized the data – what next

- What was I asking?
  - Typically: “which genes changed expression patterns when I did ____”
  - Typical contexts
    - Binary conditions: knock out, treatment, etc
    - Unordered discrete scales: multiple types of treatment or mutations, tissue types etc
    - Continuous scales: time courses, levels of treatment, etc
- Analysis methods vary with question of interest

Common questions

- Which genes changed expression patterns?
  - Statistical tests
- Which genes can be used to classify or predict the diagnostic category of the sample?
  - Machine-learning class prediction methods e.g. Support vector machines (References)
- Which genes behave similarly over time when exposed to treatment
  - Cluster analysis (Gabriel Eichler)

Selecting a subset of genes

- Question - which genes are (most) differentially expressed?

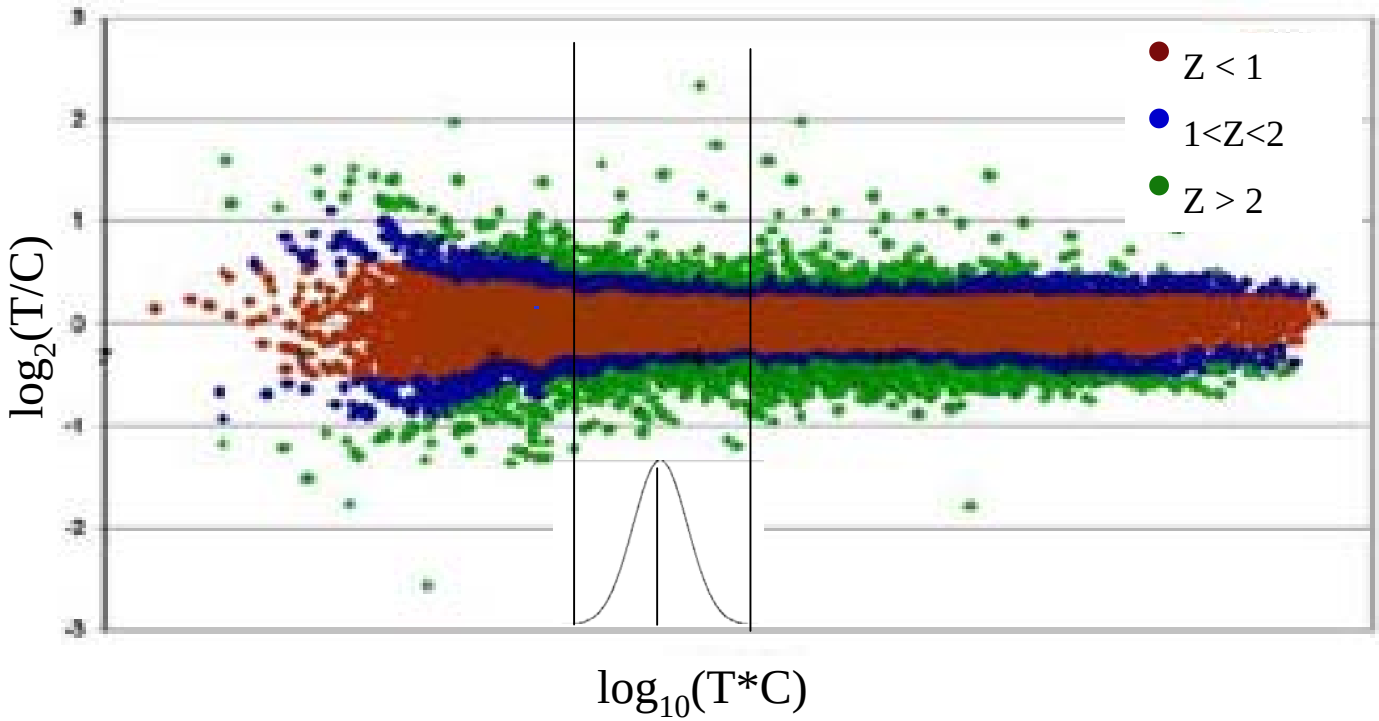
- Common Methods
  - Fold change in expression after combining replicates
    - Genes that have fold changes more than two standard deviations from the mean or pass the Z-score test
  - Statistical tests
    - Diagnostic experiments (two treatments) ***t-test***

      A t-test compares the means of the control and treatment groups
    - Multiple treatments – ***ANOVA F test***

      test the equality of three or more means at one time by using variances.

Significance – Z-scores

Intensity dependent Z-scores to identify differential Expression

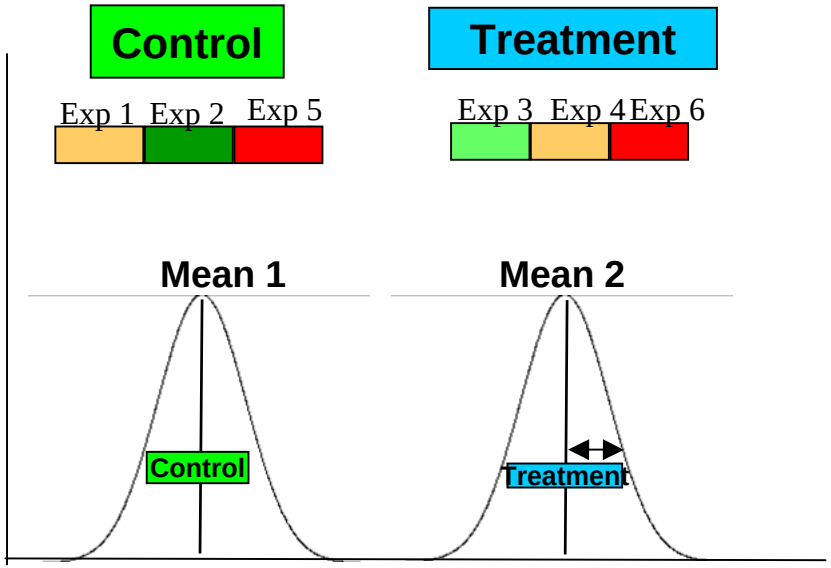


$$Z = \frac{(\log_2(T/C) - \mu)}{\sigma}$$

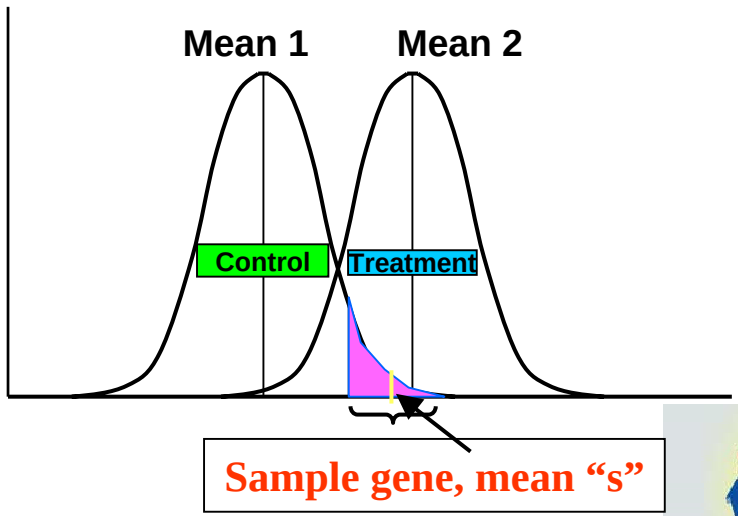
where μ - mean log(ratio)

σ - standard deviation

Statistical tests



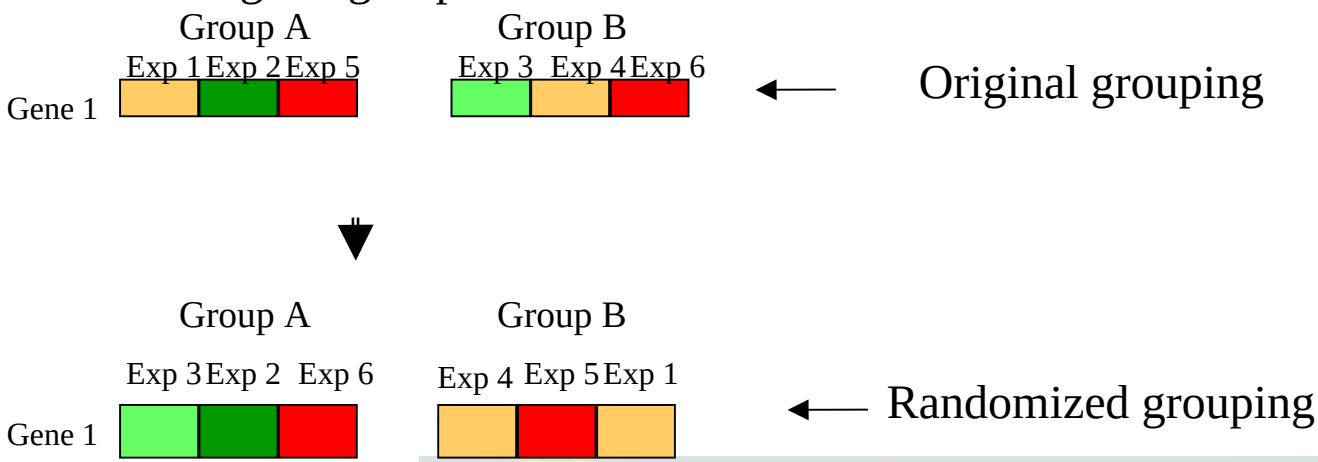
These means are definitely significantly different



Less than a 0.05 % chance that the sample with mean s came from population 1, i.e., s is significantly different from "mean 1" at the $p < 0.05$ significance level. But we cannot reject the hypothesis that the sample came from population 2.

Statistical tests – permutation tests

- Many biological variables, such as height and weight, can reasonably be assumed to approximate the normal distribution. But expression measurements? Probably not.
- Permutation tests can be used to get around the violation of the normality assumption
  - For each gene, calculate the ***t*** or ***F*** statistic
  - Randomly shuffle the values of the gene between groups A and B,such that the reshuffled groups A and B respectively have the same number of elements as the original groups A and B.



Statistical tests – permutation tests

- Compute ***t*** or ***F***-statistic for the randomized gene
- Repeat randomization *n* times
- Let *x* be the number of times ***t*** or ***F*** exceeds the absolute values of the randomized statistic for *n* randomizations.
- Then, the p-value associated with the gene = 1 – (*x*/*n*)
- The p-value of an event is a measure of the likelihood of its occurring. The lower the p-value the better
- If the calculated p-value for a gene is less than or equal to the critical p-value, the gene is considered significant.

But it may not be that simple……..

The problem of multiple testing

- Let's imagine there are 10,000 genes on a chip, AND

- None of them is differentially expressed.

- Suppose we use a statistical test for differential expression, where we consider a gene to be differentially expressed if it meets the criterion at a p-value of $p < 0.01$.

(adapted from presentation by Anja von Heydebreck, Max–Planck–Institute for Molecular Genetics, Dept. Computational Molecular Biology, Berlin, Germany **http://www.bioconductor.org/workshops/Heidelberg02/mult.pdf)**

The problem of multiple testing

- We are testing 10,000 genes, not just one!!!

- Even though none of the genes is differentially expressed, about 1% of the genes (i.e., 100 genes) will be erroneously concluded to be differentially expressed, because we have decided to “live with” a p-value of 0.01

- If only one gene were being studied, a 1% margin of error might not be a big deal, but 100 false conclusions in one study? That doesn’t sound too good.

The problem of multiple testing

- There are “tricks” we can use to reduce the severity of this problem.

- Slash p-value for each gene - each gene will be evaluated at a lower p-value.

- False Discovery Rate (FDR)- proportion of genes likely to have been wrongly identified by chance as being significant

Statistical Tests - Conclusions

- Don’t get too hung up on p-values.

- P-values should help you evaluate the strength of the evidence.

- P-values are not an absolute yardstick of significance.

Statistical significance is not necessarily the same as biological significance.

- Results from statistical tests need to be verified either in the lab (Quantitative RT-PCRs) or by comparisons to previous studies.

Time series Analysis

- Consider an experiment with 4 timepoints for each treatment.

Treatment 1: Ratio of treatment to control

Genes	Timepoint 1: Time 0			Timepoint 2: 30 min			Timepoint 3: 1 h			Timepoint 4 : 2h		
	Replicates			Replicates			Replicates			Replicates		
	1	2	3	1	2	3	1	2	3	1	2	3
Gene 1	1.2	1.5	0.8	2.7	1.5	1.3	0.6	0.9	1.2	0.3	0.4	0.4
Gene 2	-	0.8	1.05	1.2	1.7	1.4	1.8	1.8	1.6	-	-	0.9
Gene 3	0.5	1.2	-	3.7	4.3	4.1	2.8	-	2.3	1.9	2	1.4
Gene 4	0.65	0.7	0.93	0.2	0.15	-	-	-	**0.32**	-	0.76	0.87
Gene 5	-	-	0.98	-	-	-	0.65	0.87	0.54	0.87	0.85	-
Gene 6	1.67	1.1	0.87	7.32	6.43	10.54	4.54	4.32	5.3	-	2.31	2.12
Gene …												
Gene …												

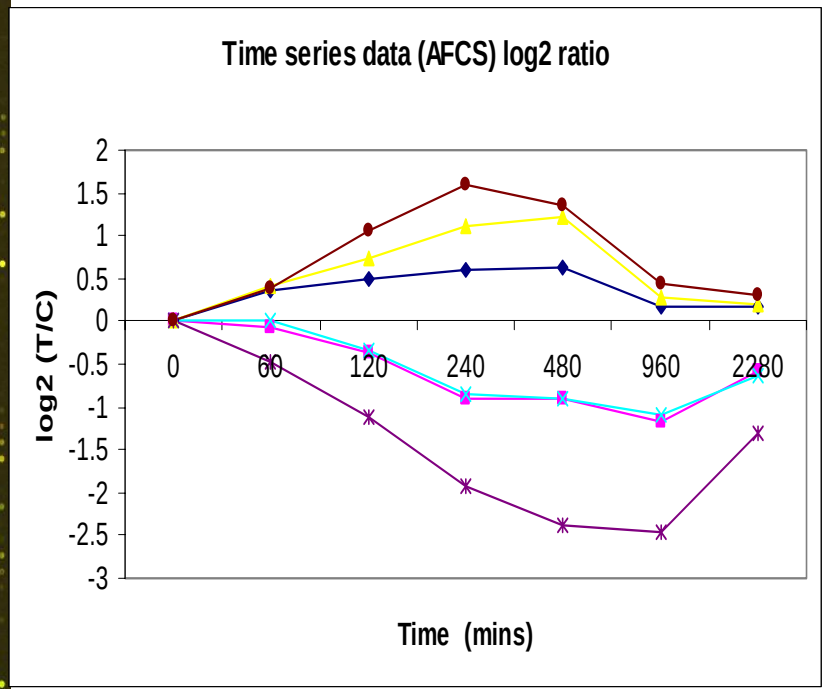
Ratio at Time 0 should be 1 in perfect circumstances

Time Series Analysis : Genes of interest

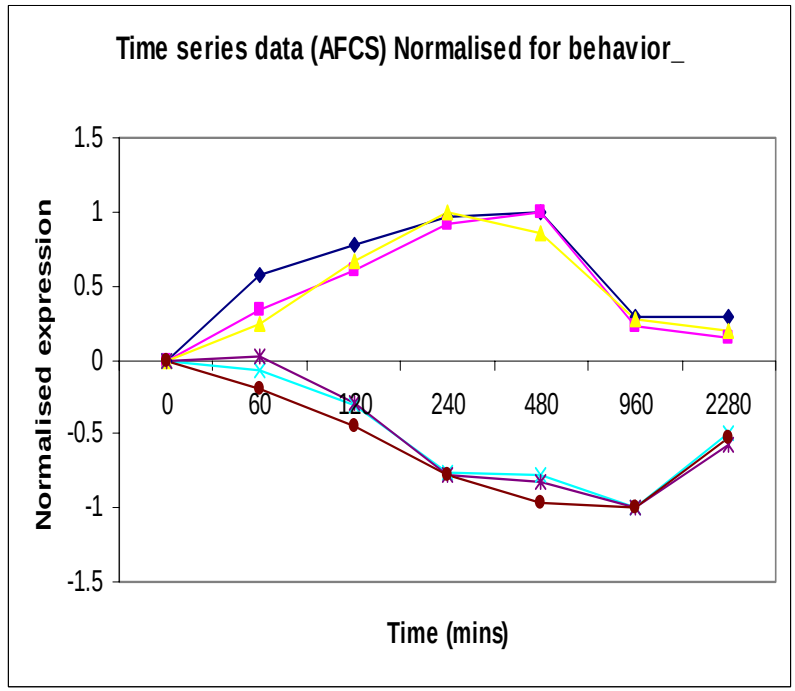
- Treatment effects
- Time effects
  - on reference or control
- Interaction between treatment and time
  - Identifying genes that are co-expressed.
    - Prediction of function.
    - Identifying a set of co-regulated genes
    - Identifying regulatory modules.

Time series Analysis - transforming data

- The time series profile has both behavior and amplitude



- The time series profile has only behavior information



- What I have covered
  - Various normalization techniques
  - Replicate filtering
  - Various statistical methods for selecting differentially expressed genes.
  - Time series data transformations
- What's next (by Gabriel Eichler)
  - Clustering algorithms
  - Concatenating time profiles from various treatments – GEDI
- Other tools
  - Supervised machine learning algorithms (suggested papers for reading)
  - Post clustering analysis (introduced in part this morning)
    - GO terms
    - Analysis of promoter elements for transcription factor binding sites

Analysis of replicates- replicate trim

The lowess-adjusted log2(*R/G*) values for two independent replicates are plotted against each other element by element. Outliers in the original data (in red) are excluded from the remainder of the data (blue) selected on the basis of a two-standard-deviation cut on the replicates.