Selecting Differentially Expressed Genes

by Jake Sauter

Motivation

- In all comparative studies (healthy vs disease, treated vs untreated, drug A vs drug B) a very important problem is to determine the genes that are **differentially expressed** (DE) in the two samples being compared
- This task is simple in principle, though becomes more complex in reality due to numerous sources of fluctuation and noise
  - For spotted cDNA arrays, non-negligible ~0.05 probability that hybridization of any spot will not reflect the presence of mRNA
  - The probability that a single spot will provide a signal even if mRNA is not present is ~0.10

Combative Production Methods

- Affymetrix technology tries to respond to the challenge of poor reliability by using a set of multiple probes to represent a gene
    - Each gene is represented by two probes, one being a perfect match (PM) and the other having a mismatch (MM) in the middle of the sequence
    - The average difference in PM and MM is taken as representative of the for the expression level of the gene
- Illumina microarrays have a large number of beads carrying the same DNA sequence, constituting technical replications (allow the estimation of variance)

Criteria

In order to assess the performance of a gene selection method, quantifiable criteria must be devised to analyze the outcome of the selection process

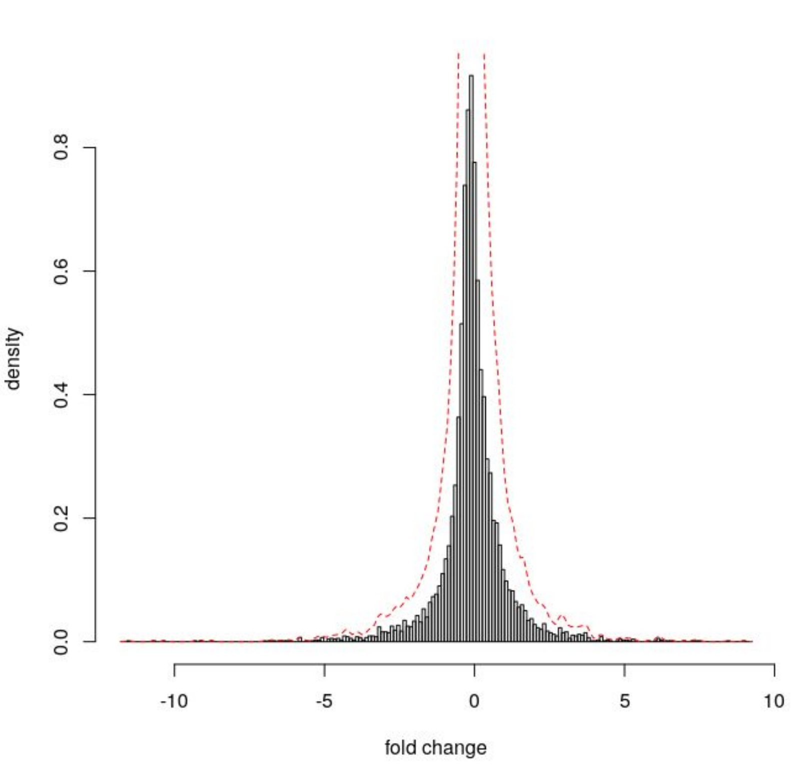
	True		
Reported	changed	unchanged	
changed	TP	FP	Positive predicted value $\frac{TP}{TP+FP}$
unchanged	FN	TN	Negative predicted value $\frac{TN}{TN+FN}$
	Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{TN+FP}$	Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

Fold Change

- **Fold Change** is the simplest and most intuitive approach to finding genes that are DE
- The log of this ratio is normally taken to provide better distribution characteristics
- A mean expression level for each condition is calculated
- Genes are selected as DE if they have a fold change greater than an arbitrary selected threshold

Fold Change

- The log of these fold change ratios can be plotted as a histogram, with the x-axis in fold change units. Selecting DE genes relates to selecting ratios on the tails of the distribution

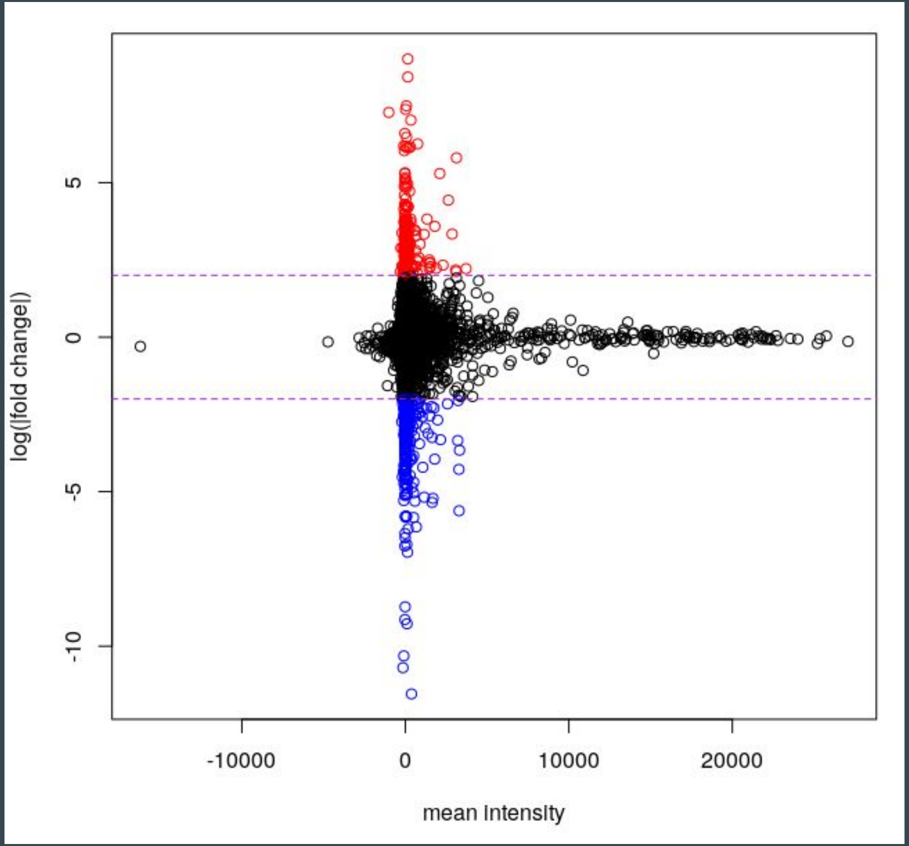


Fold Change

- Fold change is often used because it is simple and intuitive, however it has important disadvantages
  - The fold threshold is chosen arbitrarily and may often be inappropriate (no genes or non-DE genes can be selected)
- Microarrays tend to have low signal to noise ratios for genes with low expression levels (low intensity values have higher variance, high intensity values have lower variance)
  - Constant fold change threshold for all genes will introduce false positives at low expression levels (decreasing specificity) and miss true positives at high expression levels (decreasing sensitivity)

Fold Change Applied

- The log of the absolute values of the fold change ratios had to be taken due to the normalization techniques applied to the data
- 547 genes were selected as DE with a fold change threshold of 2



Unusual Ratio

- This method is superior to the fold-change method while still being simple and intuitive
- The cut-off threshold is automatically adjusted
  - Thresholds on how different the experiment/control ratio of a gene is with respect to the mean of all ratios is used instead of thresholds on the values of the ratios themselves
- No matter how many genes are up/down regulated, and no matter by how much they are regulated by, this method will always pick the genes that are affected the most

Unusual Ratio

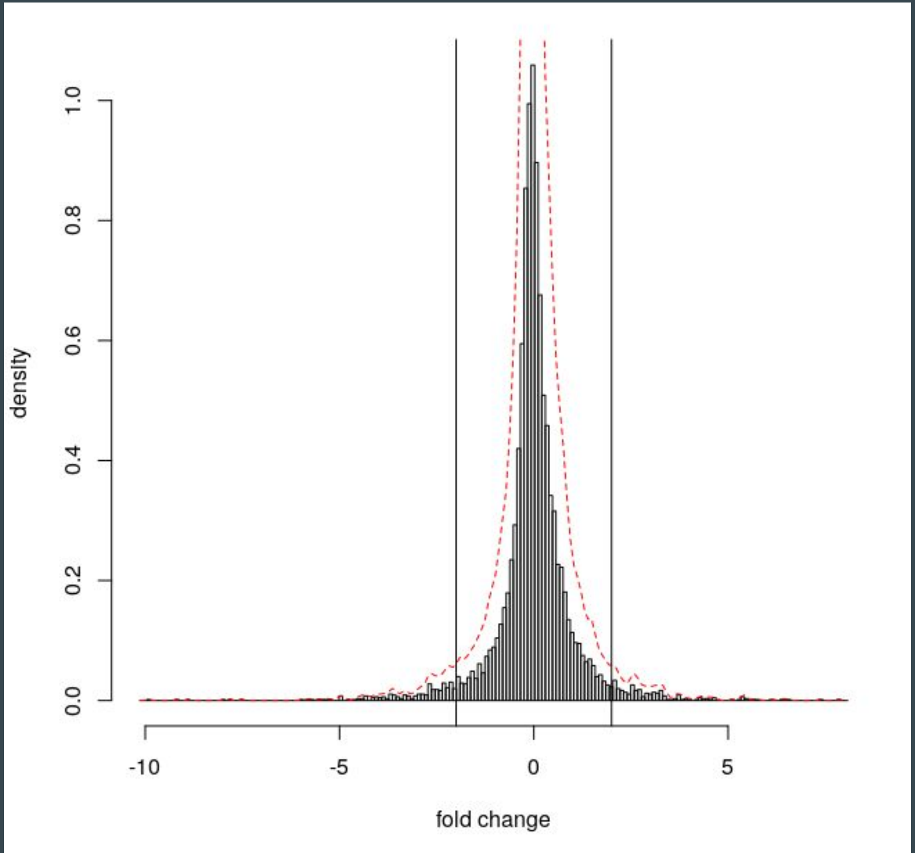
- Also a widely used tactic for selecting DE genes
- For this method, a z-transform is applied to the log ratio values
- Selects genes that are a certain distance from the mean experiment/control ratio
    - Typically this distance is taken to be $\pm 2\sigma$

Unusual Ratio

- This method still has important intrinsic drawbacks
  - Will report 5% of the genes as DE even if no DE genes are present
  - Will report 5% of the genes as DE even if many more genes are in fact DE
  - Continues to use cutoff boundaries unrelated to the high variance of low expression levels and low variance of high expression levels
- A variation of this method selects those genes for which the absolute difference in the average expression intensities is much larger than the estimated standard error computed for each gene using array replicates (if present)

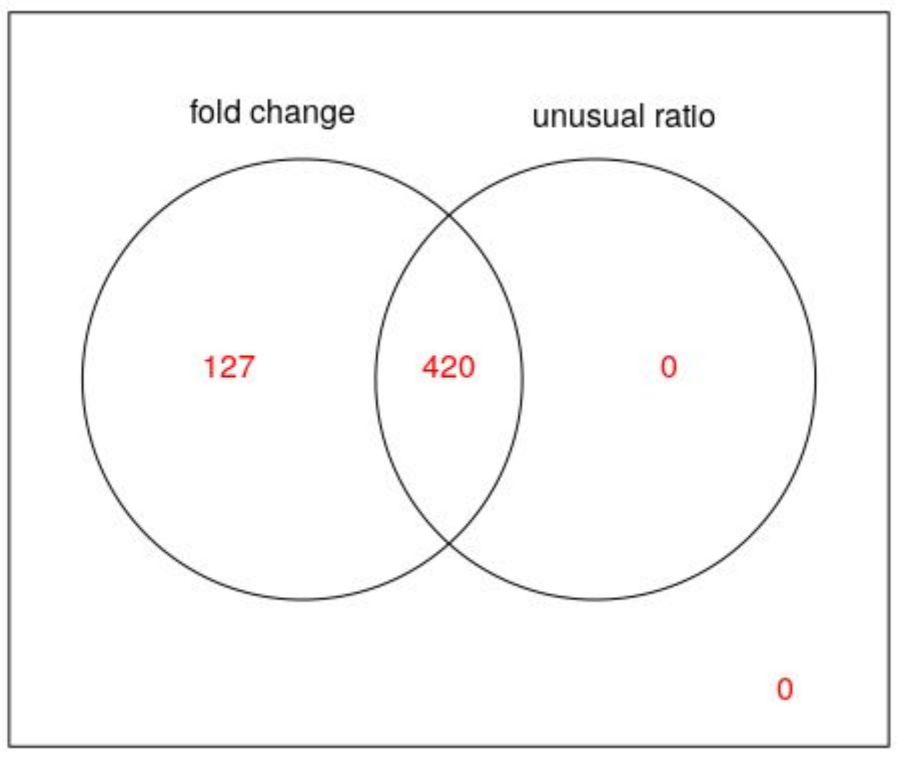
Unusual Ratio Applied

- The log of the absolute values of the ratios were z-scored before this method was performed
- 420 genes were selected as DE, being 2$\boldsymbol{\sigma}$ from the mean ratio



Fold Change and Unusual Ratio

- All genes selected by the unusual ratio method were also selected in the fold change method
- With a fold change threshold of 2 (really 4 as we have taken the base 2 log), we see that 127 more genes are selected than are in the unusual ratio method with 2σ threshold



Hypothesis Testing

- Another possible approach to selecting DE genes is to use a univariate statistical test (e.g. t-test)
- Significance level corrections for multiple comparisons must be made when doing simultaneous hypothesis tests
  - The Bonferroni and Šidák corrections for multiple comparisons have been discussed previously
  - The Holm stop-down group of methods, false discovery rate (FDR) and significance analysis of microarrays (SAM) are all suitable methods for multiple comparison corrections in the context of microarray data

Hypothesis Testing

- The drawback of hypothesis testing for finding DEs is that they tend to be conservative
  - There may just be insufficient data to reject the null hypothesis
  - Though the genes that are selected as DE are very likely to be so
- The classical hypothesis testing approach assumes that the genes are independent, which is clearly untrue in the analysis of genetic data sets
  - Combining classical hypothesis testing approaches with a re-sampling or bootstrapping approach (step-down methods or SAM) tends to make the test less conservative and take the dependencies into account

Hypothesis Testing - Gene Filtering

- As discussed, the problem with hypothesis testing is that when performing multiple tests, out significance level must change with the number of tests
    - The best way to avoid the issues that arise with multiple comparisons is to perform only as many tests are necessary
- To prevent too many tests from being performed, it is best to first filter out genes that that are array pike controls and probes that are either expressed at a very low level or exhibit little variability between samples
    - To avoid any bias being introduces, label/group information cannot be used in this filtering step

Hypothesis Testing - Gene Filtering

- Common filters ensure that
  - There are at least a few samples in the group that have significant expression values
  - The ratio of maximum / minimum intensity is at least 1.5 (the gene is variable)
- When the Golub (1999) data was filtered only for max / min intensity difference of 1.5, only 2012 of the original 7129 genes remained

Hypothesis Testing - ANOVA

- ANOVA can be used to build and explicit model of all sources of variance that affect the measurements and use the data to estimate the variance of each individual variable in the model
  - ANOVA requires a complex experimental setup so is not normally performed
- To see the complexity, the Kerr and Churchill model is as follows

$$log(y_{ijkg}) = \mu + A_i + D_j + G_g + (AD)_{ij} + (VG)_{kg} + (DG)_{jg} + \epsilon_{ijkg}$$

where the noise of all experimental interactions are modelled to produce the log ratio for gene *g* of variety *j* measure on array *i* using dye *j*

Hypothesis Testing - Noise Sampling

- A variation of ANOVA can be used to identify DE genes using spot replicates on single chips
    - Noise is estimated and confidence levels for gene regulation can be calculated
- This method modifies the Kerr-Churchill model as follows

$$logR(gs) = \mu + G(g) + \epsilon(g, s)$$

logR(gs) - log ratio for gene *g* on spot *s*

μ - average log ratio over the whole array

G(g) - a term for the differential regulation of gene *g*

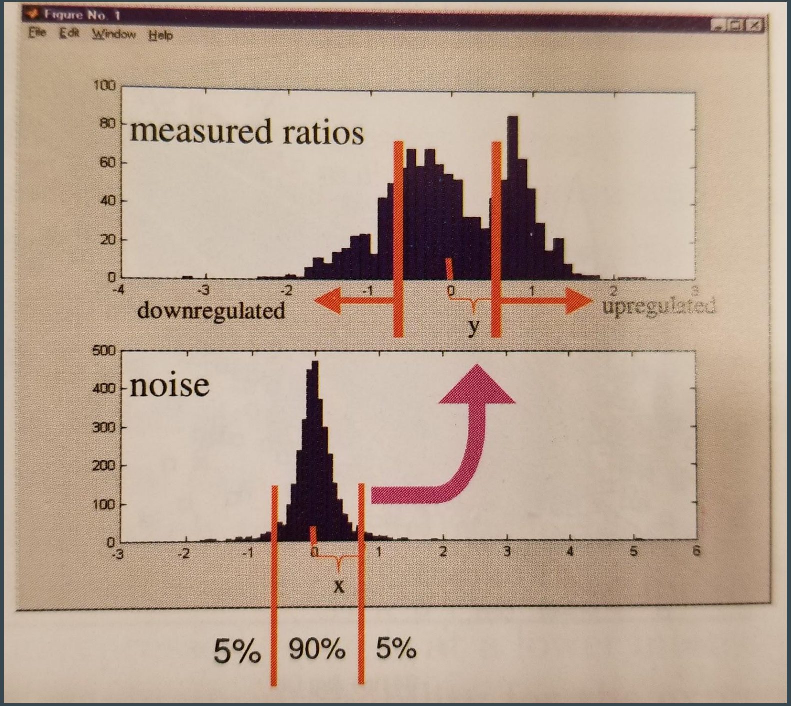
ϵ(g,s) - zero-mean noise terms

Hypothesis Testing - Noise Sampling

- In this model, estimates of the terms can be formulated to provide an estimate for the noise of each gene at each spot
    - The noise samples from each spot can be collected to yield an empirical noise distribution
    - A given confidence level can be associated with a deviation from the mean of this distribution
    - This distance on the noise distribution then can be corresponded to a distance on the measured distribution by bootstrapping
    - Dependency between intensity and variance can be taken into account by constructing several models covering the entire intensity range

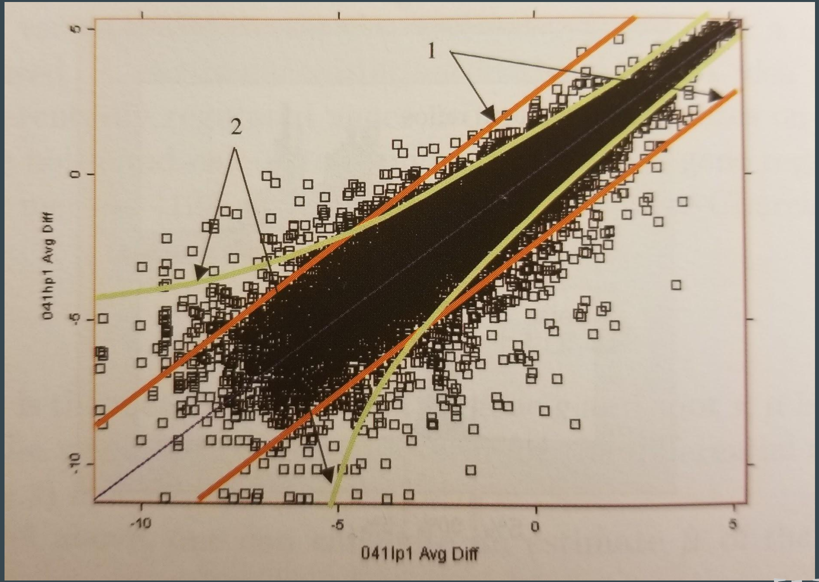
Hypothesis Testing - Noise Sampling

Visualization of correlation of noise distribution to gene expression ratio distribution



Hypothesis Testing - Noise Sampling

- The noise sampling method is advantageous as it possesses nonlinear selection boundaries that adapt automatically both to various amounts of regulation and different amount of noise for a given confidence level



[1]

(1) - Fold Change

(2) - Noise Sampling

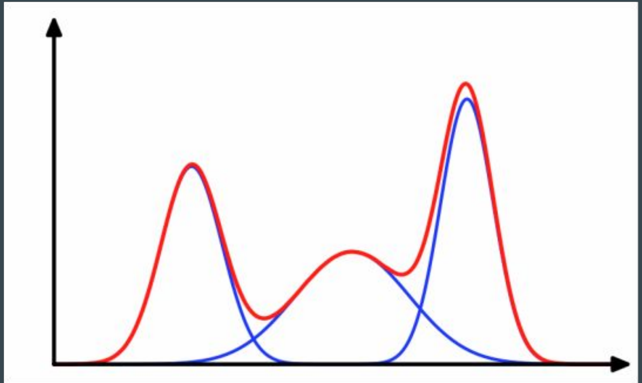
Hypothesis Testing - SAM

- SAM is software designed by Stanford specifically for applications in microarray analysis
- Uses a "relative difference" statistic d_i for gene *i*
    - Very similar to a t statistic with equal variance, except that the "gene-specific scatter" (std. dev. of the difference) s_i in the denominator is offset by a "fudge factor" s_0
    - This s_0 is an "exchangeability" factor chosen to minimize the coefficient of variation of d_i
    - This small positive exchangeability factor stabilizes d_i for genes with low expression levels
- A permutation test is used to assess significance of d_i, as well as estimate the FDR

Model-based MLE

- If p_1 , p_2 and p_3 are all a priori probabilities of a gene being expressed as up, down or not regulated; f_{up} f_{down} and $f_{unchanged}$ are all pdf's of observing value *y* ; then f_j (*y*) is a mixture model of three distributions for the probability of the observed value of the gene occuring

$$f_j(y) = p_1 \cdot f_{up_j}(y) + p_2 \cdot f_{down_j}(y) + p_3 \cdot f_{unchanged_j}(y)$$



[2]

Model Based MLE

- Then in using Bayes Theorem, we can calculate the probability of the gene in question being DE (The even E_g)

$$Pr\{E_g \mid Y_{gj} = y\} = \frac{p_1 \cdot f_{E_j}(y)}{f_j(y)}$$

Model Based MLE

- This method assumes that the model is a mixed normal pdf, so only mean and variance must be estimated for all distributions
    - These parameters can be estimated numerically using a maximum likelihood approach, which searches various combinations of parameters until the obtained equation fits the data as best as possible
- This approach has been used to provide several important facts about studies of this type
    - Any spot can provide erroneous results, thought the probability of three or more spots being erroneous is negligible
    - The intensities do seem to be normally distributed
    - The probability of a false negative is as high as 5% for any single replicate
    - The probability of a false positive is as high as 10% for any single replicate

Model Based MLE

- This approach is very general and powerful
  - The MLEs become unbiased minimum variance estimators as the sample size increases
- However this approach has many disadvantages
  - Requires solving complex nonlinear equations (computationally intensive)
  - The experimental method for isolating the distributions is complicated, and involves three a priori probabilities
  - Results become quickly unreliable as sample size decreases

References

[1] Drăghici Sorin. Statistics and Data Analysis for Microarrays: Using R and Bioconductor. Chapman and Hall, 2012.

[2] Kagan, Michael. Machine Learning: Lecture 1, CERN, 9 July 2018, indico.cern.ch/event/726959/attachments/1683504/2705968/Kagan_Lecture1.pdf.

Noise Sampling

$$logR(gs) = \mu + G(g) + \epsilon(g, s)$$

$$\hat{\mu} = \sum_{g} log(R(g, s))$$

$$\hat{G(g)} = \frac{1}{m} \sum_{g} log(R(g, s)) - \hat{\mu}$$

$$\hat{\epsilon(g, s)} = log(R(g, s)) - \hat{\mu} - \hat{G(g)}$$