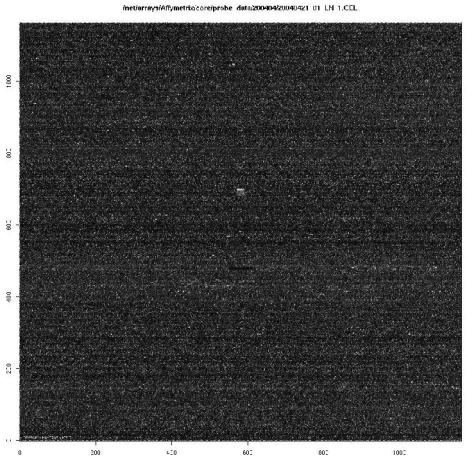
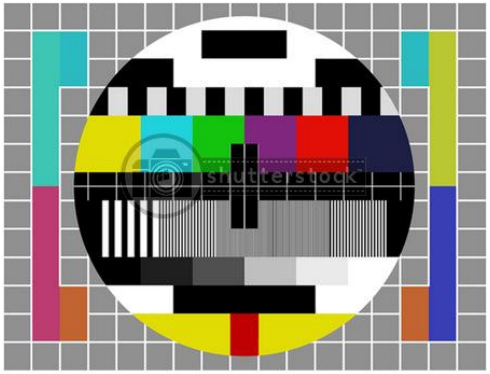
Microarray data analysis, intensive course 3.-5.9.2007.





Signal? or... No signal?

Jukka Corander,Department of Mathematics, Åbo Akademi University
Laura Elo, BTK, Turun Yliopisto

Tuesday 4.9.07

Outline:

1. Replication and estimation of expression levels
2. Two-group comparisons
3. Statistical models and experimental design
4. Microarray data mining and related approaches

1. Replication and estimation of expression levels

- Multiple copies of a gene being present on an array allows for assessment of the gene's variability.
- Hybridizing a number of arrays to the same labeled mRNA sample allows the assessment of variability across arrays (such arrays are *technical replicates*).
- Hybridizing a number of arrays to different *labeled* mRNA samples prepared from the same mRNA sample allows the assessment of variability related to labeling and sample preparation (such arrays are again *technical replicates*).
- Several mRNA samples from a number of different subjects allows assessment of biological variability (e.g. animal to animal, tissue to tissue etc). Such arrays are *biological replicates*.

Replication continued

- It is important to notice that the increase in the amount of information gained by increasing the number of replicates only holds for the particular source of variation represented by the replicates.
- For instance, increasing the number of arrays exposed for the same mRNA sample does not lead to an improvement of understanding the biological variation.
- To increase the overall precision of an experiment, it is most effective to add replicates for the source with largest level of variation.
- In most cases the largest source of variation is the biological one, which unfortunately is also the costliest one to account for due to technical issues in the sample preparation.
- See the Churchill (Nature 2002) paper for a further discussion on replication.

Technical replicates and estimation of expression levels

- Basically, the idea with technical replicates is to gain precision in the estimation of expression levels for the measured genes.
- The most basic approach to estimating the expression level is to use a normal $N(\mu,\sigma^2)$ model to describe the variation over the technical replicates for a single gene.
- The standard estimates are then the sample mean and variance.
- However, these are sensitive to deviations from the underlying normal model (e.g. skew distributions, long tails, outliers).
- Therefore, robust estimators of the central location and variation for a distribution are often to be preferred.

- Median (say M_g) is a typically used estimate for the central location of the expression level (X_{gi}) of gene g.
- Median absolute deviation (MAD) is often used to estimate standard deviation.
- $MAD_g = \text{median}\{| X_{gi} - M_g |\}$.
- These estimators are very resistant to outliers, but they also have a high variability.
- Other robust estimators generally used include trimmed means and M-estimators, such as biweight mean and standard deviation.
- However, a model-based approach to estimation offers a more systematic way of handling the expression level uncertainty (more later).

Biological replicates and estimation of expression levels

- Biological replicates represent natural variability among subjects, e.g. due to genetic diversity, environmental effects, etc.
- Such variation also contributes uncertainty to the intensity measurements associated with a gene.
- ”Averaging” across the biological replicates allows gene expression levels to be estimated with greater biological precision.
- Obviously, the replicate numbers play again a major role here.
- Biological variation can mostly be treated in a similar way as the technical replicates, both with simple estimators and statistical models.

Describing both types of variation in a statistical model

- Recall the example from the preprocessing lectures, dealing with an experiment with 10 pairs of arrays C1A,C1B,...,C10A,C10B.
- Each pair of arrays correspond to a single mRNA sample (labeled C1,...,C10) which was taken from a mouse and hybridized to two separate arrays (A and B).
- As the number of technical replicates is the same (two) for all biological replicates, the experiment is called *balanced* with respect to replication.
- Obviously, if the number of technical replicates varies across the biological replicates, the experiment is unbalanced.
- Balanced experiments have advantages over unbalanced ones from the statistical perspective.

Lets look at a simple statistical model for this type of an experiment

- Let X_{gij} denote the observed intensity of the jth technical replicate within the ith biological replicate for the gth gene.
- The indexing works as: $g = 1,...,G$; $j = 1,...,a$; $i = 1,...,n$.
- To represent both sources of variability, we may set up the following model:
- $X_{gij} = \mu_g + \alpha_{gj} + \varepsilon_{gij}$
- Here μ_g is the overall (true) mean for gene g, α_{gj} is the effect of the jth biological replicate, and ε_{gij} is the effect of the ith technical replicate.

- The variation would be further specified by stating the following assumptions about the first two moments:
- $\alpha_{gj} \sim (0,\sigma^2_{BIOL;g})$
- $\varepsilon_{gij} \sim (0,\sigma^2_{TECH;g})$
- The overall sample mean $\sum_j \sum_i X_{gij} / an$ is then a straightforward estimate for the true mean expression.
- Similarly, mean squared errors across the biological and technical replicates can be used to estimate the variation terms in the model ($\sigma^2_{BIOL;g}$, $\sigma^2_{TECH;g}$).
- However, these estimators are again sensitive to outliers, and thus, more robust alternatives are to be preferred.
- For example, median-of-medians (MOM) has been suggested as an estimator for μ_g
- Further, resampling procedures have been suggested for obtaing standard errors for this estimate.

Estimation of expression levels for multiple oligonucleotide arrays

- Li and Wong (the developers of dChip software) have suggested the following method for estimating expression levels for oligonucleotide arrays.
- Let PM_{ij} and MM_{ij} denote the observed intensities of the jth probe pair (perfect match/mismatch) within the ith array.
- The PM-MM differences are characterized by the reduced Li-Wong model as:
- $Y_{ij} = PM_{ij} - MM_{ij} = \theta_i \varphi_j + \varepsilon$
- Thus, apart from the random error, the match difference equals the product of an expression index θ_i and a probe sensitivity index φ_j.
- Statistically, we recognize this as an exponentiated form of an ANOVA model.

- The corresponding ANOVA model can be written as:
- $log(Y_{ij}) = log(\theta_i) + log(\varphi_j) + \varepsilon$
- We recognize this as a two-way model with a probe pair effect and an array effect.
- Further, the model has no intercept or interaction terms.
- The advantage of using the ANOVA framework is that more complex designs can be accounted for, using the routinely available tools.

Estimation of fold change in two-channel experiments

- Let X_{1gi} and X_{2gi} denote the log transformed and normalized intensity measurements for the channel 1 and 2, respectively.
- Assume $X_{cgi} \sim (\mu_{cg}, \sigma^2_{cg})$, $c = 1,2$; $i = 1,...,n$.
- One of the principal obejctives of two-channel array experiments is to estimate the true differential expression $\rho_g = \mu_{1g}, - \mu_{2g}$ for gene *g* and to pick out those genes that appear to be most differentially expressed.
- A straightforward estimate of this alien called the *log fold change* is provided by the mean difference $\sum_i (X_{1gi} - X_{2gi})/n$.
- Thus, an estimate of the fold change is obtained by exponentiation of the mean difference.
- Here again, several more advanced methods, such as Bayes and empirical Bayes approaches have been proposed.
- One of the reasons supporting the use of more advanced approaches is the fact that a given fold change can have a very different interpretation for a pair genes, such that one has low expression levels in both channels and the other has high levels.

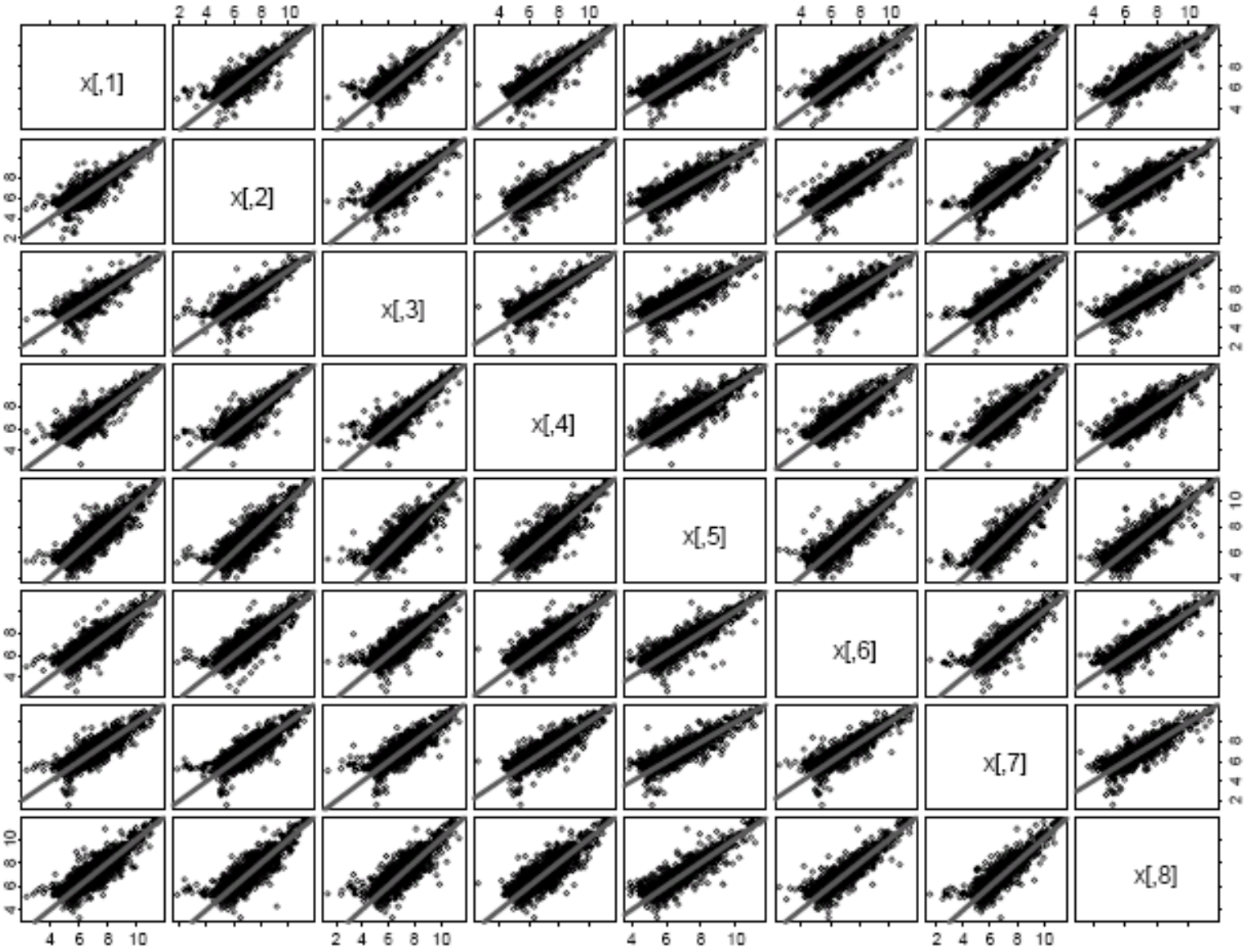
2. Two-group comparisons

- A majority of microarray experiments are comparative in nature, meaning that the purpose is to compare the expression levels of a set of genes across two or more conditions (e.g. cancer cells vs normal cells).
- In particular, one aims to identify that are differentially expressed to a considerable level.
- The simplest approach is to consider each gene in isolation with respect to this issue.
- The simplest situation where this can be done is the comparison of two groups.
- We consider more complicated situations and approaches later on.

An example

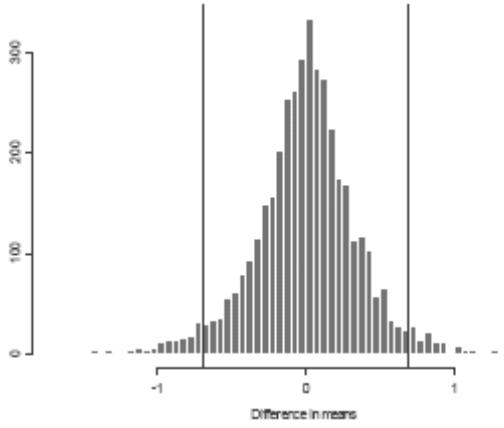
- Two groups of mice, referred to as Group 1 and 2, respectively.
- Each group has four mice and the members of the group have received a group-specific treatment.
- After treatment a mRNA sample was extracted from the liver of each animal.
- Each sample was hybridized with an array containing 4077 genes.
- The following picture shows the pairwise scatterplots of the log-transformed and normalized expression levels of the eight mice.

Pairwise expression scatterplots



Fold changes

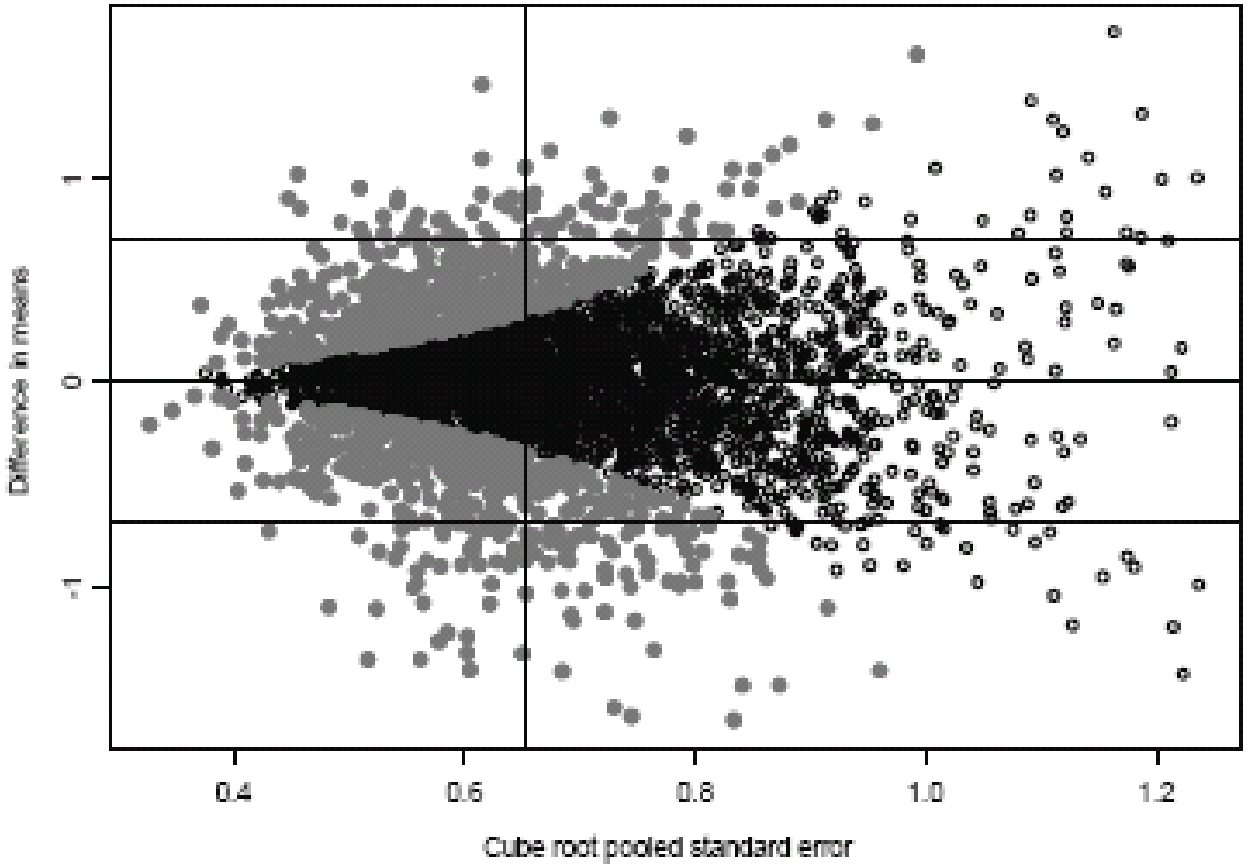
- Early approaches to expression level comparison simply declared a gene differentially expressed, if its fold increase or decrease exceeded a specified cut-off point (such as 2 or 5).
- The picture below shows a histogram of the fold changes with vertical lines indicating the cut-off value equal to two, which leads to 119 significantly upregulated and 144 significantly downregulated genes.



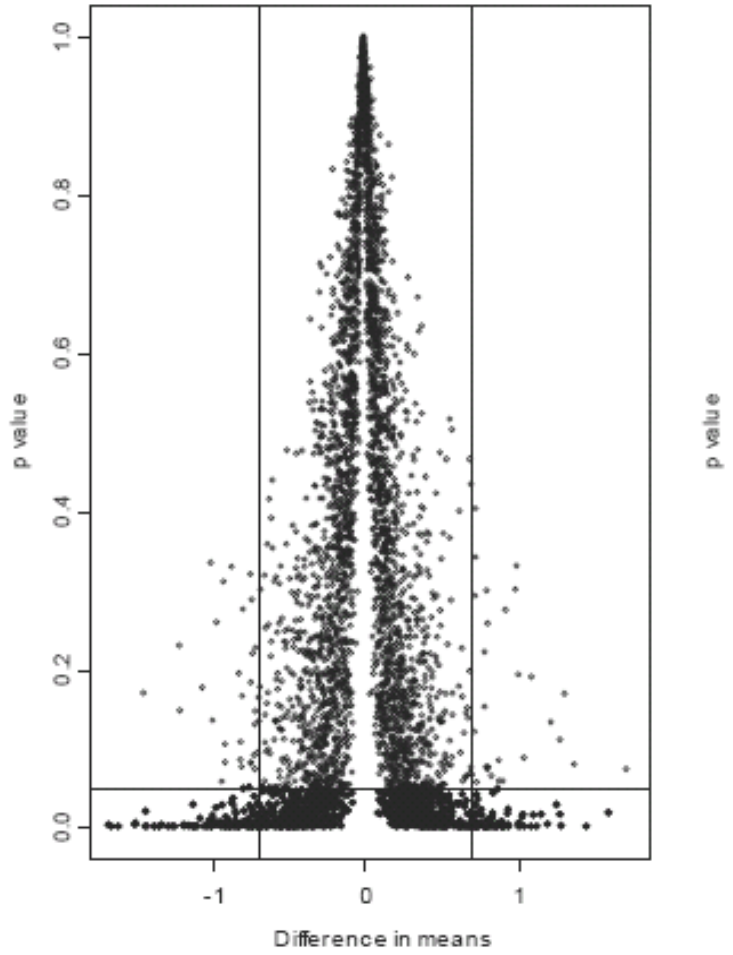
- Use of the fold changes as such to screen for significant genes is subject to criticism.
- Remember that the used expression estimates (such as means) are just a statistical representation of the unknown expression values, and thus subject to variability.
- Genes with high variability have quite high probability of having a high fold change, and thus may look deceptively interesting.
- It may be that a gene has 10-fold change and yet not be biologically significant due to its high variability.
- On the contrary, a gene with twofold change may be highly significant if it associated with a low level of variability.

- Perhaps the most basic approach to taking the variability into account is the two sample *t* test:
- $T_e = |m_1 - m_2|/s_p^*$, where
- m_1 and m_2 are the group sample means and s_p^* an adjusted square root of the pooled estimate of variance for the two samples.
- In the example data, 998 genes out of 4077 are significantly differentially expressed according to a *t* test at 5% level.
- This test has tendency to ignore genes that have large difference in means, if they also have high variances.
- This is reasonable in principle, but the test appears to focus too much on genes with small variances in the current setting where the variances are estimated from very small samples (which makes them to be less well-estimated).
- Notice that the form of the test corresponds to a signal-to-noise ratio, the numerator being the 'signal' and the denominator being the 'noise'.

Behavior of the t test components for the example data (open circles are non-significant genes)



Behavior of the t test components for the example data (continued)

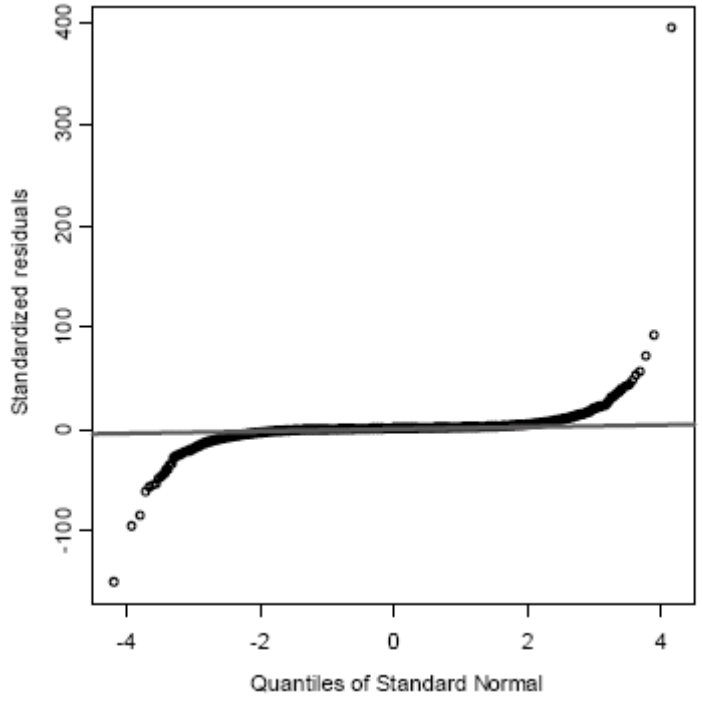


- The *t* test is based on assuming normality of the data with equal variances for the two groups, and as these assumptions may be unjustified quite easily, the behavior of the test is affected accordingly.
- If the underlying distribution has longer tails than normal distribution, the denominator of the statistic is inflated.
- In this case the test will have both lower false positive rate and higher false negative rate than expected, because it will be harder to achieve significance.
- If the normality assumption is judged reasonable, but the variances are deemed different in the two groups, then another form of *t* test, Welch's test, is available.
- In the example data, 872 genes out of 4077 are significantly differentially expressed according to Welch's test at 5% level, which is somewhat less than the set of genes fished out by the *t* test.

Diagnostic checks

- Residuals ($r_{ij} = x_{ij} - m_j$) provide a basis for checking distributional assumptions, such as normality (here m_j refers to group median instead of the mean, to avoid outlier effects).
- Sorted and plotted against the quantiles of a normal distribution (so called normal probability plot) they provide immediate visual means for judging the reliability of the assumptions.
- Microarrays have typically tapering away at the edges, indicating long-tailedness of the distribution.
- Several tests are also available for assessing the normality assumption, such as Shapiro-Wilk and variants of the Kolmogorov-Smirnov test.
- However, the graphics are often quite sufficient for this purpose.

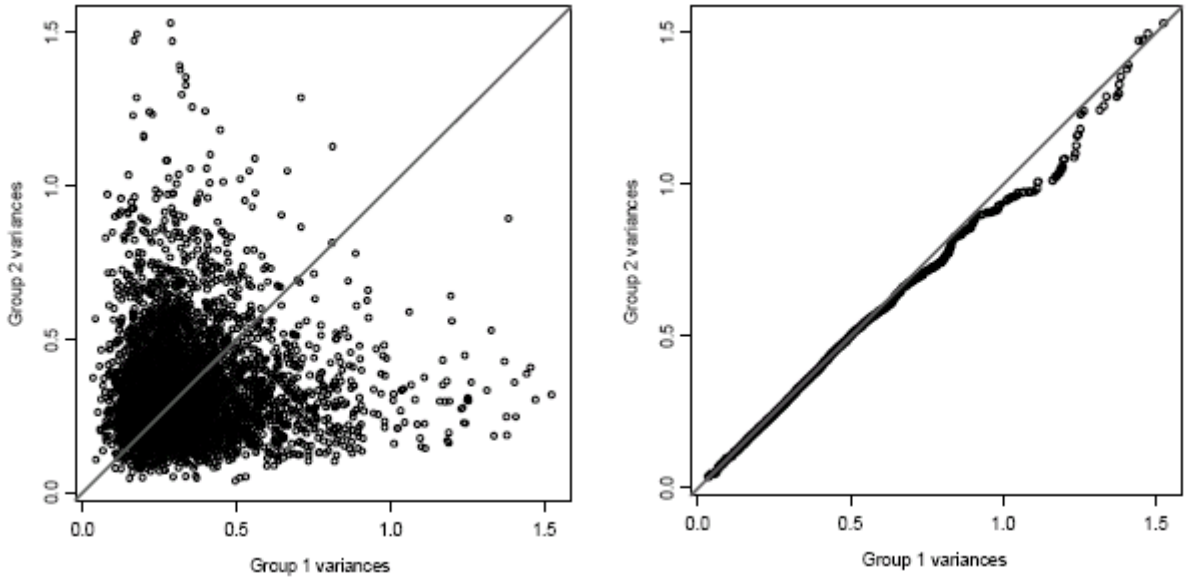
Normal probability plot for the example data (straight line is the identity line).



Diagnostic checks continued

- There exist formal tests to check the assumption of equal variances for two samples (such as Bartlett's and Levene's tests), however, they are not considered useful for typical microarray studies due to the small sample sizes.
- Here again graphics are more helpful (see the next page).
- *t* tests can also be made more robust to outliers by replacing the sample statistics by robust versions of them, such as biweight or trimmed estimators.
- This approach is typically able to reduce the high false negative rate of the test under the influence of outliers, by modifying the elements of the "signal-to-noise" ratio.

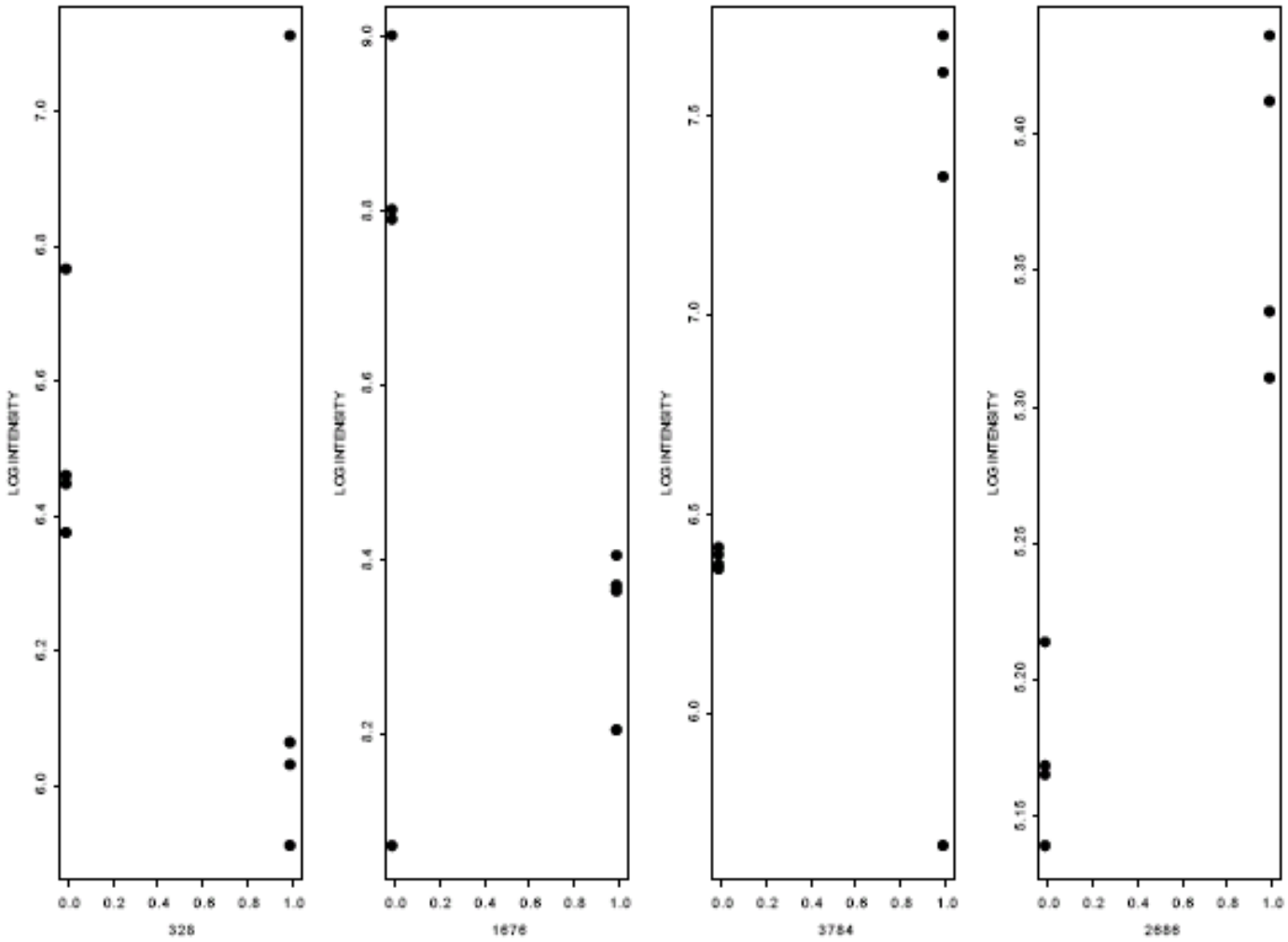
Variance plots for the example data (straight line is the identity line).



We see that in this particular case the variances for the two groups are quite equal over the gene set.

- A robust version of the *t* test for the example data gives quite different results, leading to 228 significantly upregulated and 224 significantly downregulated genes.
- This is quite much less than the 998 genes indicated by the earlier test.
- However, even if the robust test yields less significant genes due to some loss of power, there are also numerous cases where it results in a significant difference, whereas the ordinary *t* test yields a nonsignificant result.
- This is primarily due to the resistance to outliers, which is illustrated on the next page, where single outliers hamper the *t* test .

Figure 7.5: Log intensities for the control group and the treatment group for four genes, the first three of which are significant by the robust *t* test but not significant by the *t* test, the fourth of which is not significant by the robust *t* test but significant by the robust *t* test.



Note: there is a typo in the fig. caption, word ’robust’ should be erased from the last line.

Randomization tests

- The test methods based on randomization procedures are used in the two-group comparison according to the following steps:
- Choose a test statistic T and calculate its value T_{obs} for the observed data.
- Permute randomly the pooled set of observations from the two groups.
- Assign the n_1 first observations to one group and the remaining observations (n_2) to another group.
- Calculate the value of the test statistic, say T_p for this grouping of the data.
- Repeat the permutation and calculate finally the empirical probability for $T_p > T_{obs}$ (i.e. # permutations with T_p exceeding T_{obs}, divided by the total number of permutations).
- This probability serves as the p-value under the null hypothesis of no difference among the two groups.
- Note that the number of possible distinct permutations may be quite small.

- A randomized version of the ordinary *t* test for the example data gives again quite different results, as compared to the earlier ones.
- There are 35 ways of splitting the data randomly (two groups of four, i.e. 8 choose 4 divided by 2).
- By setting one permutation exceeding the observed test statistic value to be the cut-off, we have a 5.7% significance limit for a two-sided test.
- This yields 1384 significantly differentially expressed genes, which is much more than earlier.
- It is important to notice that while the randomized test removes the necessity of specifying the underlying distribution, the test is robust to outliers only if the test statistic itself is resistant to them.
- For example, a randomization test based on means is not robust.

Nonparametric tests

- There exists a group of test methods that are called nonparametric or distribution-free tests, a well-known example being the Mann-Whitney-Wilcoxon test.
- This test uses the rank sum for one of the groups calculated from a pooled and ordered sample, as the distribution of the sum is tractable if the two samples emanate from the same unknown underlying distribution.
- Unfortunately, the power to detect differences is rather small for nonparametric tests, given the typically small sample sizes (per gene!) in microarray studies.

Test multiplicity

- A test-based analysis of differences among groups necessitates the calculation of a very large number of tests in an ordinary microarray study.
- Thus, by the very principles of the hypothesis testing, a large number of false positives is expected to be discovered (α multiplied by #genes).
- The statistical literature contains a large body of work related to the multiple hypothesis testing problem.
- Old approaches include corrections to p-values, such as the Bonferroni and Sidak's methods.
- For the example data, Bonferroni correction yields 12 significant genes.
- Later methods include a sequential approach, such as that pioneereed by Sture Holm.
- Unfortunately, all these approaches are suboptimal due to their large corrections that reduce the power excessively (the same 12 genes are found by the Holm-Bonferroni method).

A pragmatic approach to dealing with test multiplicity

- One failure of an ‘objective’ statistical testing approach is that it fails to take into account the specific purposes of the microarray study at hand.
- Thus, taking a pragmatic approach, one can judge how much the p-values should be adjusted depending on what the goals of the study.
- A side note: incidentally, the Bayesian approach to statistics also tells the modeler to focus on the purpose to be achieved by modeling the data...
- For example, in a screening study one could be willing to accept a fairly large number of false positives in order to improve the chances of capturing truly differentially expressed genes.
- In another study, where the purpose is to identify the most promising genes for further consideration, e.g. in a knock-out experiment, one would have to only choose a handful of genes due to limited resources for processing them.

- Both goals can be reached by specifying subjectively a number H of genes that one is willing to accept as positive discoveries.
- By ranking all the genes according to some criterion, e.g. p-value, one can then choose the H best ranking ones, given that they all would be significant by a threshold deemed suitable (otherwise one can still drop those among the H genes that are not even significant).
- However, the intricacies of the testing procedures outlined above should be taken as a warning signal that the naive testing approach is really suboptimal for this “large p small n” type of a problem.
- Statistical modelling approach is definitely to be preferred!

False discovery rate approach

- Given the difficulties with the control of many *p*-values, an approach to controlling false discovery rate (FDR) has been suggested instead.
- The description of the FDR approaches is somewhat involved in the course book, so lets take a look at the Wikipedia definition.
- Positive false discovery rate (pFDR) is discussed in detail in Storey (Annals Stat 2003), link to the paper is available at the Wikipedia page.
- He shows also that pFDR has a clear and intuitive Bayesian interpretation, namely that it equals the posterior probability of a false positive given that the test statistic is significant.
- Surely, hundreds of different flip-flop methods exist for attacking the significance problem for two-group comparisons for microarrays. However, it’s time to leave this issue and proceed further...

3. Statistical models and experimental design

- After the initial statistically crude approaches used to investigating microarray data in the early years of the technology, an increasing number of works have exploited general ideas from experimental design and linear models.
- Here we consider some basic designs to illustrate the general principles.
- Remember the paper by Churchill (Nature Genetics 2002), which provides an excellent discussion.

- Consider a simple comparative array experiment whose objective is to investigate how/which genes express differentially across a single factor *V*.
- This factor can represent things such as treatments, time, tissues etc.
- Some authors call these instances “varieties”.
- The indexing works here as: $g = 1,...,G$ (# genes); $j = 1,...,J$ (# arrays); $i = 1,...,I$ (# varieties).
- Let Y_{gij} be a suitably transformed and normalized expression level measurement for gene *g* in array *j* assigned to variety *i*.

- The simplest approach is to model the data for *each gene separately,* using a linear model such as
- $Y_{gij} = \mu_g + V_{ig} + \varepsilon_{gij}$
- Here μ_g is the average signal for gene g, V_{ig} is the additional signal due to the effect of the ith variety on gene g, and ε_{gij} is the error not accounted for by the other terms in the model.
- The traditional assumption is to set $\varepsilon_{gij} \sim iid\ N(0,\sigma^2_g)$.
- The model can be fitted through a least squares procedure and F statistics can be used to screen for significant differences among the varieties.
- Note that the F statistic can be seen as a simple extension of the earlier discussed *t* statistic.
- Therefore, all the headaches we discussed earlier apply to this situation as well, except that working around them is even trickier.
- I.e. the problems with outliers and the varying levels of expression need still to be accounted for.

An example

- An experiment with 9 mice was conducted to study the response to a particular drug.
- The mice were all treated with the drug, and after 1,2 and 3 hours, 3 mice were on each occasion randomly chosen and a sample of mRNA was extracted from their livers.
- In addition, there were 3 control mice to represent the 0 hour situation.
- A dozen arrays were each exposed to the sample from a single mouse.
- Using F test for each gene separately, 334 genes out of 2004 were found to be significantly differentially expressed across the varieties.
- Using the Bonferroni correction, only 2 genes remained significant.

Treating genes simultaneously

- Lets now look at modeling the genes simultaneously, instead of treating them in isolation from each other.
- In the simple comparative study discussed above, there are three different *effects* or *factors* that can influence the expression levels.
- These are varieties (*V*), arrays (*A*) and genes (*G*).
- It is reasonable to formulate a model that describes the relationship between the expression levels Y_{gij} and the three factors and their possible interactions.
- Lets look at the potential terms to be included in such a model.

Main effects

- An array effect (A), would account for overall differences in expression level measurements after the effects of other model factors are removed.
- If the normalization step is successful, an array effect should be fairly small.
- A gene effect (G), would account for differences among the average expression levels among the considered gene set (e.g. due to variation in hybridization efficiency, variation in the natural expression level, etc).
- A variety effect (V), would account for differences in expression level measurements, if some varieties are associated with substantially higher or lower level overall levels than others.

Two-factor interaction effects

- A variety-gene interaction effect (VG), would account for how a gene expresses differentially across the varieties.
- Given a particular gene *g*, if any of the $(VG)_{gi}$ terms is larger than the other relative to the underlying variability, it means that the particular variety is inducing a higher level of expression than the other varieties.
- Contrasts among the $(VG)_{gi}$ terms are of primary interest in comparative microarray studies.
- An array-gene interaction effect (AG), would account for the variability of a spot across the arrays averaged over all the spots. This effect would be observed if the concentration or amount of DNA assigned to microarrays varies from array to array.
- A variety-array interaction effect (VA), would account for variability across the varieties for arrays. However, this effect is not estimable when an array contains only a single variety.

The simplest additive linear model

- This simultaneous model includes the factors V,A,G,VG:
- $Y_{gij} = \mu + V_i + A_{j(i)} + G_g + (VG)_{gi} + \varepsilon_{gij}$
- Here μ represents the average signal across the whole experiment.
- The traditional assumption is to set $\varepsilon_{gij} \sim iid\ N(0,\sigma^2)$.
- However, for a typical array data set, it is possible that none of the assumptions of normality, independence and homoscedasticity (equi-variance) holds particularly well.

Reasons for failing model assumptions I

- Nonnormality:
- Empirical evidence seems to indicate that it is reasonable to assume the error distribution being symmetric and somewhat normal-like (bell shaped in the middle).
- However, the problem is related to the tails, which are heavier than normal and also contain outliers.
- In addition, signal saturation causes truncation effects for high expression levels.

Reasons for failing model assumptions II

- Lack of independence:
- Genes are rarely expressed in isolation but along biological pathways.
- Therefore, the assumption of independent expression levels for different genes is unrealistic.
- As a result, the test patterns will be correlated.
- However, it is quite difficult to model the gene correlation structure in advance and the typical data sizes do not easily allow one to infer the correlations from the data.
- Recently, Bayesian methods have been developed as attempts to solving this learning problem.

Reasons for failing model assumptions III

- Heteroscedasticity:
- In some experiments the genes appear to have equal levels of variations after suitable transformations, in which case the assumption $\varepsilon_{gij} \sim N(0,\sigma^2)$ is quite acceptable.
- However, in most cases the genes exhibiting high expression levels also tend to exhibit high levels of variation.
- Therefore, the assumption concerning the error variances could be modified to $\varepsilon_{gij} \sim (0,\sigma^2_g)$, i.e. gene-specific variance.
- In addition, in some cases it is necessary to consider the variation separately for different varieties, i.e. $\varepsilon_{gij} \sim (0,\sigma^2_{gi})$.

Fitting the model in two stages

- In the previous description of the effects involved in a microarray experiment, it is natural to divide the effects into two groups:
- Global effects, i.e. those not involving specific genes.
- Gene-specific effects.
- In statistical terms these effects are orthogonal to each other, which suggests a two-stage strategy to fitting the model to ease the computational burden.
- It is possible to fit first a normalization model:
- $Y_{gij} = \mu + V_i + A_{j(i)} + \delta_{gij}$
- Notice that there are no gene-specific effects in this model.
- The error terms could be assumed to behave as $\delta_{gij} \sim (0,\sigma^2_0)$.

- In the second stage, the residuals R_{gij}, from the normalization model are regarded as expression measurements that have been treated for a number of effects.
- These values are used as inputs to the gene-specific second stage model:
- $R_{gij} = G_g + (VG)_{gi} + \varepsilon_{gij}$
- Here, again, the error terms may be assumed to behave according to $\varepsilon_{gij} \sim iid\ \mathrm{N}(0,\sigma^2)$..
- The two-stage strategy will be subject to some bias due to the fact that the residuals are slightly correlated to each other, even if the effects in the two distinct groups are orthogonal.

Experimental design issues

- Consider a two-channel experiment, where two varieties (I_1 and I_2) are to be compared.
- Assume that two arrays (A_1 and A_2) are available for the purpose.
- Let us consider possible experimental designs and their properties.

- For this design, array specific effects are confounded with variety effects.
- This means that if a gene is differentially expressed in A_1 versus A_2, it is not possible to know whether attribute it to array or attribute it to variety.

Design 1	Array A_1	Array A_2
Channel R	I_1	I_2
Channel G	I_1	I_2

- For this design, potential dye-specific effects are confounded with variety effects.
- It is advisable to avoid such experimental designs.

Design 2	Array A_1	Array A_2
Channel R	I_1	I_1
Channel G	I_2	I_2

- For this design, the dyes assigned to the two varieties are switched in the second array.
- It is called a dye-swap design and it allows both for dye-specific effects and array-specific effects in the linear model.

Design 3	Array A_1	Array A_2
Channel R	I_1	I_2
Channel G	I_2	I_1

The additive linear model for the dye-swap design

- This model can be written as:
- $Y_{gij} = \mu + V_i + A_{j(i)} + D_k + G_g + (VG)_{gi} + (AG)_{gj} + (DG)_{gk} + \varepsilon_{gij}$
- In the two-stage approach the model can be specified as:.
- $Y_{gij} = \mu + V_i + A_{j(i)} + D_k + \delta_{gij1}$
- Here the additional term is due to the dye contrast.
- The corresponding gene model is:
- $R_{gij} = G_g + (VG)_{gi} + (AG)_{gj} + (DG)_{gk} + \varepsilon_{gij}$

- In the cases where biological replicates are used, it is advisable to do the dye-swap design by applying the swap to the replicates, as shown below.

Design 3*	Array A_1	Array A_2	Array A_3	Array A_4
Channel R	I_{11}	I_{21}	I_{12}	I_{22}
Channel G	I_{21}	I_{11}	I_{22}	I_{12}

Example of a single-channel experiment

- Consider a single-channel experiment, where four varieties (e.g. different drug treatments I_1, I_2, I_3 and I_4) are to be compared.
- Each treatment is to be given to four animals.
- Given that a single array is used per animal, they can be labeled as A_{ij}, where i is the variety and j is the array replicate index.
- Assume now that the lab facility has the capacity to process at most four arrays per day.

- Below is one possible experimental design.
- The problem with this design is that, if there is a day effect (and such effects really exist!), the treatment effect will be confounded with it.

Day 1	$A_{11}, A_{12}, A_{13}, A_{14}$
Day 2	$A_{21}, A_{22}, A_{23}, A_{24}$
Day 3	$A_{31}, A_{32}, A_{33}, A_{34}$
Day 4	$A_{41}, A_{42}, A_{43}, A_{44}$

- A preferable experimental design is shown below.
- For this design it is possible to estimate day effects and adjust treatment-gene effects accordingly.

Day 1	$A_{11}, A_{21}, A_{31}, A_{41}$
Day 2	$A_{12}, A_{22}, A_{32}, A_{42}$
Day 3	$A_{13}, A_{23}, A_{33}, A_{43}$
Day 4	$A_{14}, A_{24}, A_{34}, A_{44}$

4. Microarray data mining and related approahces

- As we have seen, the arrays typically contain a large number of genes.
- It is of considerable biological interest to find patterns of expression among sets of genes.
- Also, it is informative to try to link expression patterns to other data sources describing genes.
- Models describing dependencies among genes using networks have attained a lot of interest.
- All these goals are made difficult by the fact that microarrays represent ”large *p* small *n*” problems.

Pattern discovery

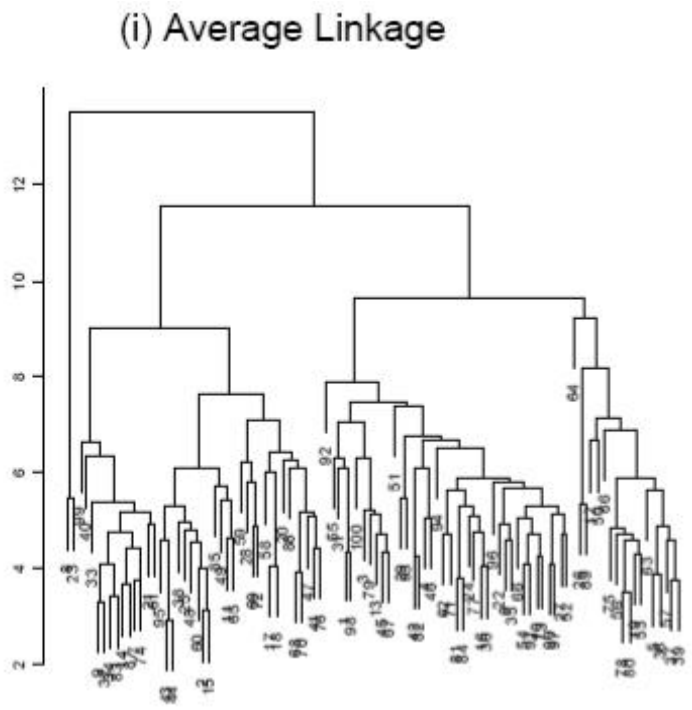
- Expression pattern discovery can be done by exploiting various standard statistical techniques.
- It can be purely algorithmic or based on statistical models.
- Popular algorithmic tools include PCA, SOM, k-means clustering and hierarchical clustering.
- An advantage of these methods is that they are widely available in software packages and computationally fairly rapid (although even these methods start easily coughing when confronted with the large gene sets present in oligo arrays!).

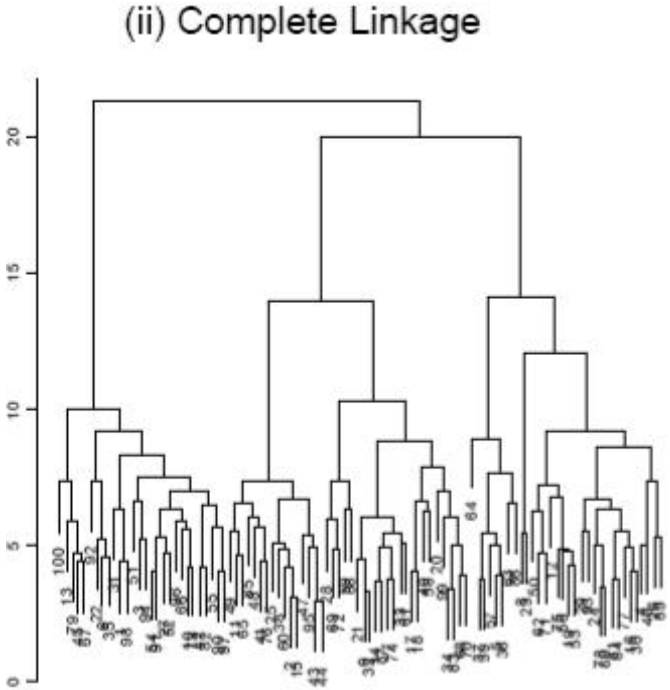
- The disadvantage of these methods is that making firm conclusions and judging statistical uncertainty can be quite difficult (e.g. determining the number of clusters).
- Therefore, a large number of model-based methods have been suggested in the literature for the same purposes.
- A disadvantage of such methods is that they are typically much more computationally intensive.
- Also, they require sometimes more statistical expertise to be applied appropriately.
- Nevertheless, they can be really rewarding by offering possibilities to do directly statistical inference about different aspects of biological interest.

Hierarchical clustering methods

- Such methods are based on calculating first some measure of distance between the genes.
- Many alternative distances are available and it is not typically to decide what should be an appropriate one.
- Therefore, a general practice is to do the analysis with a variety of distances (and clustering algorithms) to see which results seem to make sense from the biological perspective.
- One problem is that different results can make different biological sense equall well, so which one should be chosen?
- Unfortunately, a tempting option is to choose the result one happens to like most due to some reasons.
- Different generally used hierarchical clustering algorithms include single, average and complete linkage, Ward and centroid clustering.

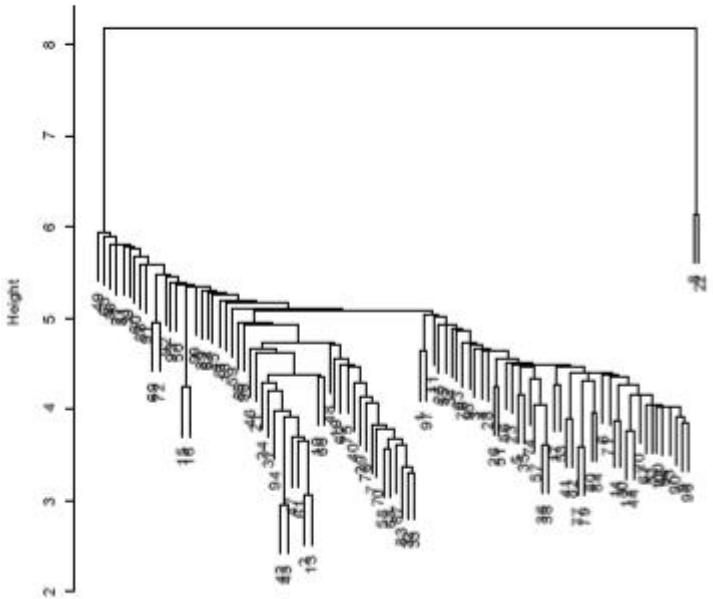
Example with two tumor tissues I



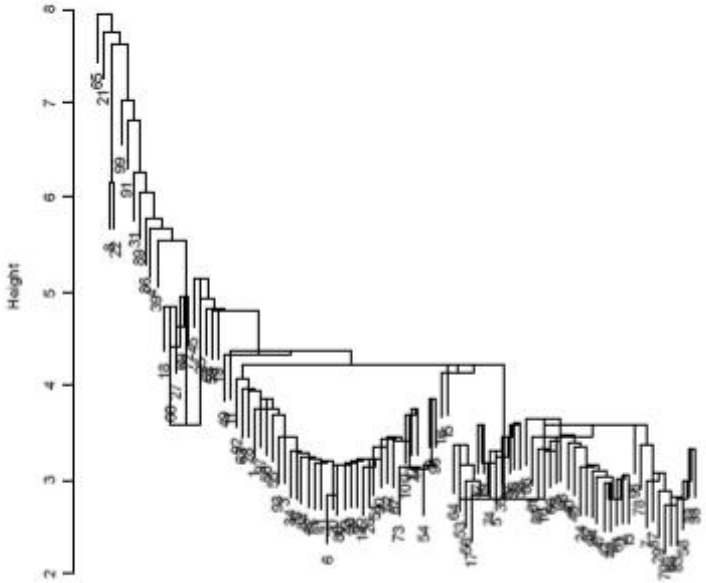


Example with two tumor tissues II

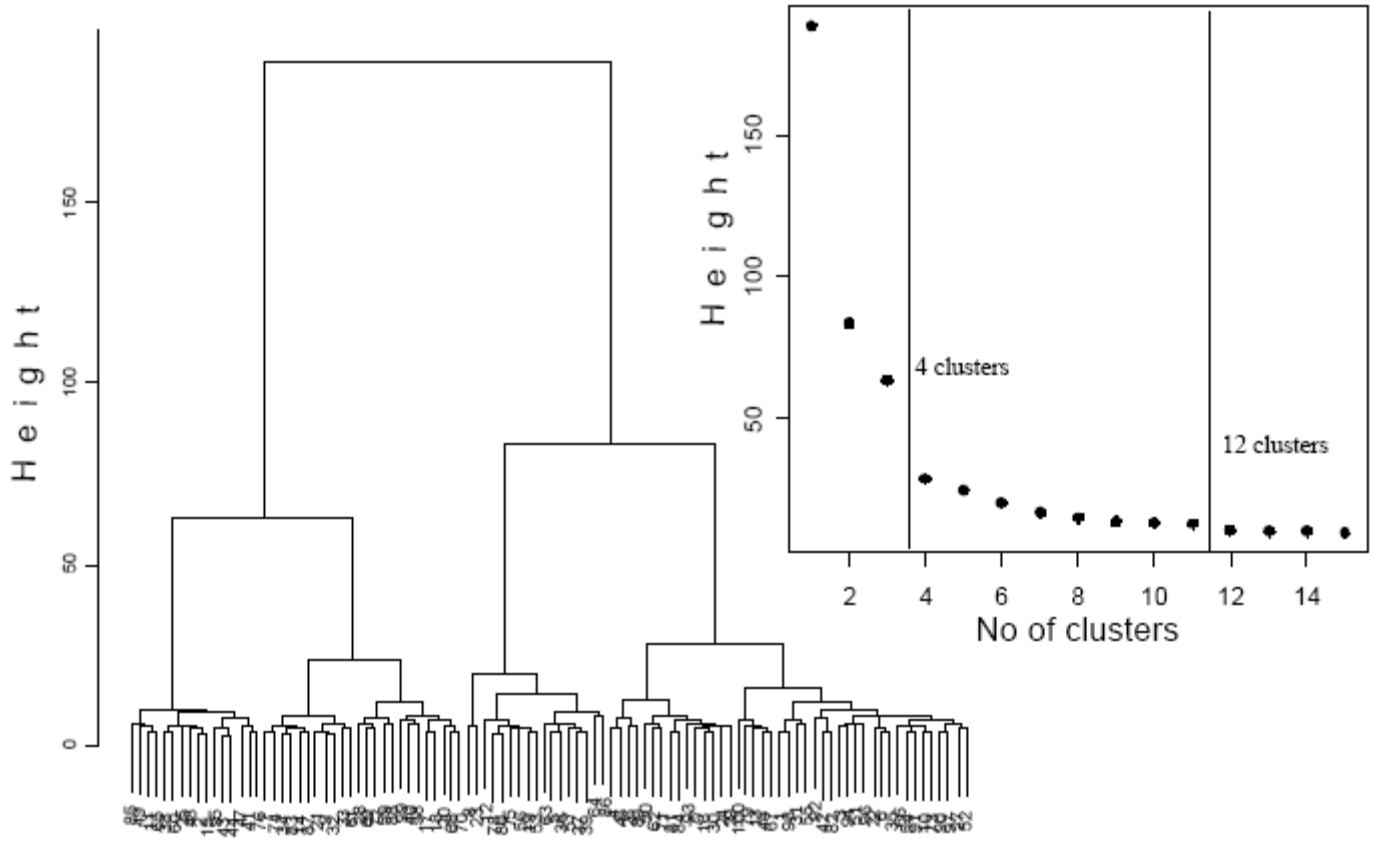
(iii) Single Linkage



(iv) Centroid Linkage



Example with two tumor tissues III

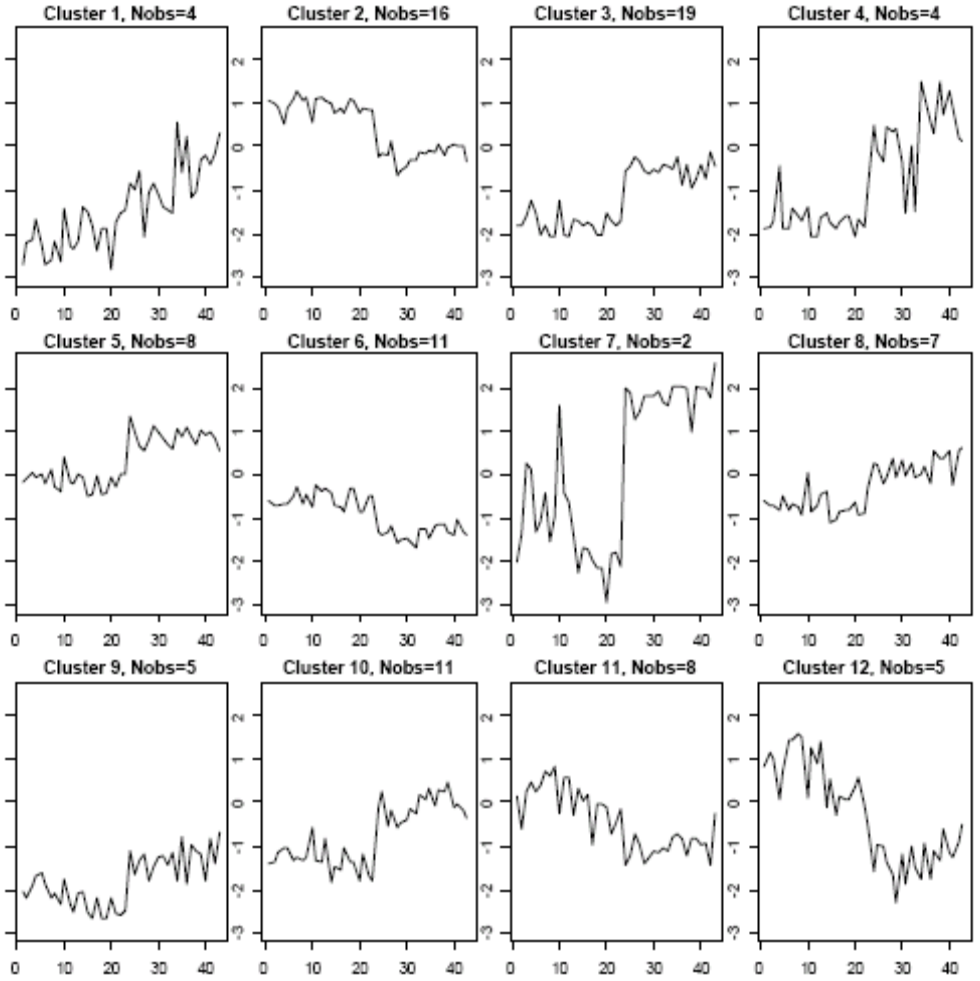


Ward's method was used for this tree.

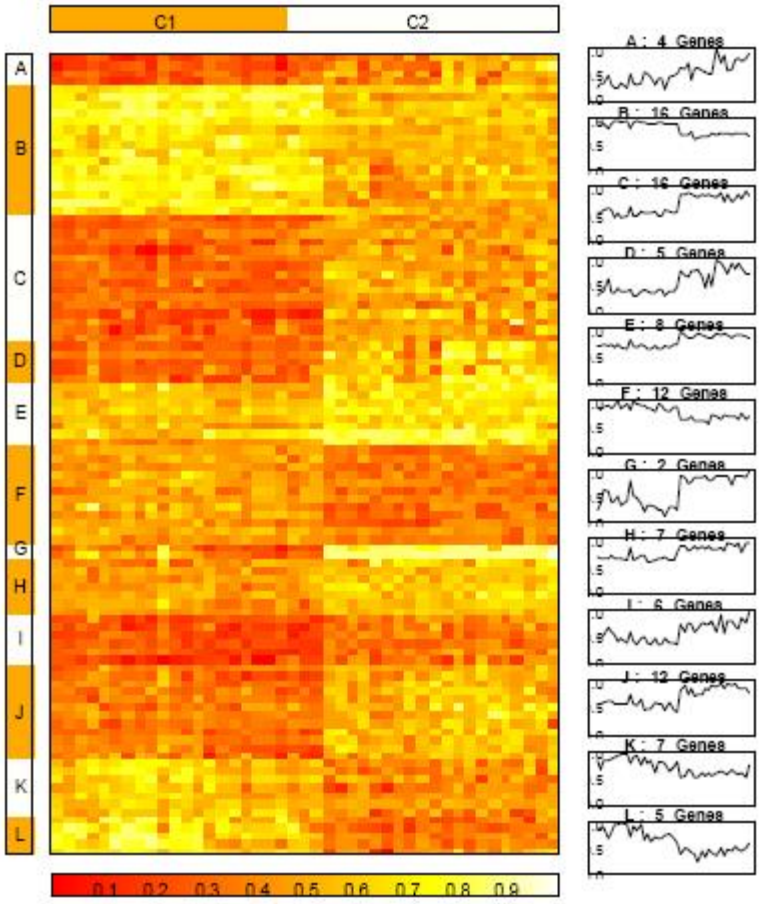
Moral of the story: Not always easy to judge the number of clusters to be chosen from a hierarchical clustering method!

Morover, the same challenge is present in the algorithmic data partitioning methods (e.g. k-means).

Visual ways of looking into gene clusters I



Visual ways of looking into gene clusters II



C1 & C2 represent samples, horizontal boxes clusters