The model based approach to estimation of expression variation

Notation: We have *k* genes and *n* samples divided into *G* groups. The genes are indexed by *i*, the groups by *g*, and the samples within a group are indexed by *j*. The number of samples in group *g* is n_g. We let y_{igj} be the log transformed measured gene expression for gene *i* in sample *j* within group *g*, α_{ig} is an expression level dependent on the group *g* to which the sample belongs, and β_{gj} represents the amount of mRNA present in sample.

A natural model is to have

$$y_{igj} = \alpha_{ig} + \beta_{gj} + \varepsilon_{igj}, \tag{1.1}$$

where ε_{igj} is an error term with mean zero. The variance of the error term ε_{igj} depends on the gene *i* and is also allowed to depend on the group label. We denote the variance by σ_{ig}^2 for gene *i* and group *g*. The error terms are assumed to be independent. Our interest is to estimate the variances σ_{ig}^2 (the intra-group variation) and the differences in gene expression level between groups (inter-group) as measured through $\delta_{ig} = \alpha_{ig} - \bar{\alpha}_{i\bullet}, g = 1,\ldots,G,$ where $\bar{\alpha}_{i\bullet} = (\alpha_{i1} + \alpha_{i2} + \ldots + \alpha_{iG})/G,$ and to find the genes with the smallest values of the variances and inter-group differences. The case of one group only, $G = 1$, is simple in that we only need to estimate the intra-group variances σ_{ig}^2. In typical applications we will, however, have more than one group.

Estimation of the intra-group variances

We first discuss the estimation of the intra-group variances σ_{ig}^2. We base the estimation on moment equations so as to obtain unbiased estimates. To obtain terms that do not depend on the gene and sample levels we subtract a gene average $\bar{y}_{ig\bullet}$ (average over the samples in the group *g*)

and a sample average $\bar{y}_{\bullet gj}$ (average over the genes) and add the average over the genes and the samples in the group $\bar{y}_{\bullet g\bullet}$ to the measured value y_{igj}. Denoting the new terms by r_{igj} we calculate the sample variances s_{ig}^2 of these within each group and use these to estimate the variances. Precisely, we define

$$r_{igj} = y_{igj} - \bar{y}_{ig\bullet} - \bar{y}_{\bullet gj} + \bar{y}_{\bullet g\bullet} \text{ and } s_{ig}^2 = \frac{\sum_j r_{igj}^2}{(n_g - 1)\left(1 - \frac{2}{k}\right)}. \quad (1.2)$$

The sample variance s_{ig}^2 has mean

$$\text{mean}(s_{ig}^2) = \sigma_{ig}^2 + \frac{1}{k(k-2)} \sum_{v=1}^{k} \sigma_{vg}^2$$

From this we see that if $k > 2$ we can identify the variances and we obtain an estimate $\hat{\sigma}_{ig}^2$ with the correct mean σ_{ig}^2 by defining

$$\hat{\sigma}_{ig}^2 = s_{ig}^2 - \frac{1}{k(k-1)} \sum_{v=1}^{k} s_{vg}^2 . \quad (1.3)$$

An alternative to the moment based estimation of the variances is to use the full distribution of s_{ig}^2 r_{igj} (assuming normality of the error terms) to estimate the variances. This is the so-called REML methodology. In practice we see very little difference in the two sets of estimates. The distributional properties of the estimates (1.3) are not simple. However, the variance can be calculated so that we can attach error bars to the estimates. In the special case of normal distributed error terms a lengthy calculation shows that the variance of the estimate $\hat{\sigma}_{ig}^2$ is

$$\frac{2n_g^2}{(n_g-1)^3\left(1-\frac{2}{k}\right)^2}\left\{\left(1-\frac{2}{k(k-1)}\right)\tau_{ig}^2+\frac{1}{k^2(k-1)^2}\sum_{m=1}^{k}\tau_{mg}^2 - \frac{2k(k-1)-1}{k^2(k-1)^2}\sum_{m\neq i}\omega_{img}^2+\frac{1}{k^2(k-1)^2}\sum_{m=1}^{k}\sum_{v\notin\{i,m\}}\omega_{mvg}^2\right\}, \tag{1.4}$$

where

$$\tau_{ig} = \frac{n_g-1}{n_g}\left\{\left(1-\frac{2}{k}\right)\sigma_{ig}^2+\frac{1}{k^2}\sum_{v=1}^{k}\sigma_{vg}^2\right\},\ i=1,\ldots k,$$

$$\omega_{img} = -\frac{n_g-1}{n_g k}\left(\sigma_{ig}^2+\sigma_{mg}^2-\frac{1}{k}\sum_{v=1}^{k}\sigma_{vg}^2\right),\ i,m=1,\ldots k.$$

If it is believed that the variances σ_{ig}^2 do not depend on the group *g* a common estimate for each gene can be obtained as

$$\hat{\sigma}_i^2=\sum_{g=1}^{G}(n_g-1)\hat{\sigma}_{ig}^2/(n-G) \tag{1.5}$$

Estimation of the inter-group variation

We next study the differences in expression level α_{ig} across the groups. Let $z_{ig}=\bar{y}_{ig\bullet}$ be the average of the measured gene expressions for gene *i* in group *g* and let $\theta_g=\bar{\beta}_{g\bullet}$ be the average sample level in group *g*. Then z_{ig} has mean $\alpha_{ig}+\theta_g$ and variance σ_{ig}^2/n_g. All the information we have on α_{ig} resides in z_{ig}, but we have the problem that the unknown sample level θ_g is

confounded with the average expression level of all the genes $\bar{\alpha}_{\bullet g}$ in group *g*. Thus we can estimate $\theta_g + \bar{\alpha}_{\bullet g}$, but not θ_g alone, unless we make some further assumption. We will here use the way the genes were selected, that is, genes were selected from microarray data to show a small variation across all the samples. We therefore make the assumption that θ can be estimated so as to minimize the variation in $z_{ig} - \theta_g$. This assumption is equivalent to the assumption that the average expression level of all the genes $\bar{\alpha}_{\bullet g}$ is independent of the group *g*. With this assumption the differences in expression level $\delta_{ig} = \alpha_{ig} - \bar{\alpha}_{i\bullet}$ is naturally estimated by

$$d_{ig} = z_{ig} - \bar{z}_{i\cdot} - \bar{z}_{\cdot g} + \bar{z}_{\cdot\cdot},$$

having mean $\delta_i(g) - \bar{\delta}_\bullet$ and with the above assumption implying that $\bar{\delta}_\bullet = 0$. The variance of d_{ig} is

$$\frac{kG(k-2)(G-2)}{(kG)^2}\frac{\sigma_{ig}^2}{n_g} + \frac{k(k-2)}{(kG)^2}\sum_{w=1}^{G}\frac{\sigma_{iw}^2}{n_w} + \frac{G(G-2)}{(kG)^2}\sum_{v=1}^{k}\frac{\sigma_{vg}^2}{n_g} + \frac{1}{(kG)^2}\sum_{l=1}^{k}\sum_{w=1}^{G}\frac{\sigma_{lw}^2}{n_w},$$

which can be used for making a confidence interval for $\delta_i(g) - \bar{\delta}_\bullet$.

We have now characterized the variation in gene *i* through the estimates d_{ig} and $\hat{\sigma}_{ig}^2$, $g = 1,\ldots,G$. How do we combine these into one measure for the quality of gene *i* as a control gene? In a typical application of gene *i* as a control gene one measurement of gene *i* will be made and used in conjunction with measurements of a number of genes looking for differentiable

expression between the sample groups for these genes. The influence of gene *i* on the measurement of the differences between the groups for the gene of interest is given by the values $z_{ig} - \theta_g - \alpha_i$, $g = 1,\ldots,G$. As an example if we have two groups and $(z_{i1} - \theta_1 - \alpha_i) - (z_{i2} - \theta_2 - \alpha_i) = 0.3$ (on the logaritmic scale) then all the estimated foldchanges for the genes of interest will be off by 1.35. We do not measure $z_{ig} - \theta_g - \alpha_i$, but estimate this quantity through d_{ig}. It is, however, not the value of d_{ig} that is of interest to us, but rather the value of $z_{ig} - \theta_g - \alpha_i$ in a future experiment. So what we do is to use d_{ig} and $\sigma_i^2(g)$ from the present experiment to evaluate the distribution of $z_{ig} - \theta_g - \alpha_i$ in a future experiment, and then pick a typical value from this distribution as our “stability” measure. In order to make this calculation we assume that α_{ig} is a random value from a normal distribution with mean α_i and variance γ^2. If we imagine that we know the value of $z_{ig} - \theta_g - \alpha_i$ the distribution of a future value $\alpha_{ig} - \alpha_i$ is normal with mean $\gamma^2(z_{ig} - \theta_g - \alpha_i)/(\gamma^2 + \sigma_{ig}^2/n_g)$ and variance $\gamma^2(\sigma_{ig}^2/n_g)/(\gamma^2 + \sigma_{ig}^2/n_g)$. This implies that the distribution of a new measurement of $z_{ig} - \theta_g - \alpha_i$ is normal with

$$\text{mean} = \frac{\gamma^2(z_{ig} - \theta_g - \alpha_i)}{\gamma^2 + \sigma_{ig}^2/n_g} \text{ and variance} = \sigma_{ig}^2/n_g + \frac{\gamma^2\sigma_{ig}^2/n_g}{\gamma^2 + \sigma_{ig}^2/n_g}. \quad (1.6)$$

This distribution reflects our knowledge of what to expect in the future.

Since values far away from zero are critical to the use of the gene as a housekeeping gene we suggest using the absolute value of the mean plus the standard deviation of this distribution as a stability measure. Denoting the stability measure ρ_{ig} we find

$$\rho_{ig} = \frac{\hat{\gamma}^2 \left| d_{ig} \right|}{\hat{\gamma}^2 + \hat{\sigma}_{ig}^2 / n_g} + \sqrt{\hat{\sigma}_{ig}^2 / n_g + \frac{\hat{\gamma}^2 \hat{\sigma}_{ig}^2 / n_g}{\hat{\gamma}^2 + \hat{\sigma}_{ig}^2 / n_g}}, \tag{1.7}$$

where we have replaced $z_{ig} - \theta_g - \alpha_i$ by its estimate d_{ig}, replaced σ_{ig}^2 by its estimate and replaced γ^2 by an estimate $\hat{\gamma}^2$. We finally combine ρ_{ig}, $g = 1,\ldots,m$, into one value for gene *i* by taking the average

$$\rho_i = \sum_{g=1}^{G} \rho_{ig} / G . \tag{1.8}$$

The estimate $\hat{\gamma}^2$ is taken to be

$$\hat{\gamma}^2 = \frac{1}{(k-1)(G-1)} \sum_{i=1}^{k} \sum_{g=1}^{G} d_{ig}^2 - \frac{1}{kG} \sum_{i=1}^{k} \sum_{g=1}^{G} \hat{\sigma}_{ig}^2 / n_g ,$$

if this is positive and zero otherwise.

Average control gene

An important question is whether it is better to use the average of a set of housekeeping genes instead of a single one. If we take the average of the genes in a set *A* we no longer look at the distribution of a future value of $z_{ig} - \theta_g - \alpha_i$, but instead the distribution of a future value of the average $\frac{1}{|A|} \sum_{i \in A} \left(z_{ig} - \theta_g - \alpha_i \right)$, where $|A|$ is the number of elements in *A*. We obtain the mean and variance of this average directly from (1.6). When replacing $z_{ig} - \theta_g - \alpha_i$ by its estimate d_{ig} we

need to take into account that θ_g has been replaced by an estimate, in particular $\sum_i d_{ig} = 0$. A variance calculation shows that this can be corrected for by multiplying the mean in (1.6) by $\sqrt{k/(k-|A|)}$. Thus arguing as in (1.7) we use

$$\rho_{Ag} = \left| \frac{\sqrt{k}}{|A|\sqrt{k-|A|}} \sum_{i\in A} \frac{\hat{\gamma}^2 d_{ig}}{\hat{\gamma}^2 + \hat{\sigma}_{ig}^2/n_g} \right| + \sqrt{\frac{1}{|A|^2} \sum_{i\in A} \left(\hat{\sigma}_{ig}^2/n_g + \frac{\hat{\gamma}^2 \hat{\sigma}_{ig}^2/n_g}{\hat{\gamma}^2 + \hat{\sigma}_{ig}^2/n_g} \right)},$$

$$\rho_A = \frac{1}{|A|} \sum_{g=1}^{G} \rho_{Ag} \tag{1.10}$$

as our stability measure.

This will depend very much on the values of d_{ig} for $i \in A$. In the case $|A| = 2$ if d_{ig} has the same sign for the two genes using the average will typically not be an improvement. It seems most important to use the average in those cases where there are no genes with $d_{ig} \approx 0$, and then to use two genes with opposite signs of d_{ig}.

The pair-wise comparison approach by Vandesompele et al. (2002)

Notation: We have *k* genes and *n* samples. The genes are indexed by *i* and the samples are indexed by *j*. We let y_{ij} be the log transformed measured gene expression for gene *i* in sample *j*. Vandesompele et al. consider all the samples as belonging to one group and take as their starting point the sample variances $s^2(i_1,i_2)$ of differences $y_{i_1 j} - y_{i_2 j}, j = 1,\ldots,n$, between two genes. They then define a stability measure M_i^k as the average of $s(i,r)$ over *r* among the *k* genes. Finally, for each $l = k,k-1,\ldots,3$ they sequentially exclude the gene *i* with the highest value of M_i^l, until only two genes are left.

It is clear that the main differences between the model based approach and the approach of Vandesompele et al. are the inclusion of the groups in the former and a direct estimation of variances instead of a pair-wise comparison method.

If we consider the model based approach with one group only, then $s^2(i_1,i_2)$ is an estimate of $\sigma_{i_1}^2 + \sigma_{i_2}^2$ and the ranking of the genes obtained from $\hat{\sigma}_i^2$ (1.3) and from M_i^k will be comparable. In practice we saw minor differences even for genes with a low variance, this is most likely due to the compound nature of M_i^k. When comparing the rankings from $\hat{\sigma}_i^2$ and from the full pair-wise comparison sequential method of Vandesompele et al. the differences are bigger, although still for genes with low variances. For 4 of the 5 data sets considered in Vandesompele et al.(9) the three “best” genes selected by their method do not contain the gene with the lowest value of $\hat{\sigma}_i^2$, and in one case (pool) the two genes with the lowest values of $\hat{\sigma}_i^2$ are not included.

To evaluate the difference between our direct estimation of the intra-group variances and the sequential pair-wise comparison approach of Vandesompele et al we have performed simulations in

the simplistic situation of one group only. We have considered the situation with k=10 genes and n=8 samples and different possibilities for the variation in the 10 variances σ_i^2. We use the simulation to evaluate the probability P_{II} that the top 2 genes determined by the method used are the two genes with the lowest values of the true variances σ_i^2, and to determine the probability P_I that the top 2 genes include only one of the two genes with the lowest values of the true variances. In the first run we let $\sigma_1^2 = \sigma_2^2 = 1$ and $\sigma_3^2 = \sigma_4^2 = \cdots = \sigma_{10}^2 = \tau^2$. When $\tau^2 = 2$ we find that the Vandesomple et. al approach gives (P_{II}, P_I) = (0.10, 0.53) and our direct estimation gives (P_{II}, P_I) = (0.16, 0.60). We see here that the latter method is clearly superior. When taking instead $\tau^2 = 4$ we get for the Vandesomple et al approach (P_{II}, P_I) = (0.34, 0.48) and for the direct method (P_{II}, P_I) = (0.47, 0.48). In the second run we take $\sigma_i^2 = 1 + \tau^2(i-1), i = 1, \cdots, 10$. When τ^2 is chosen so that $\frac{\sigma_{10}^2}{\sigma_1^2} = 10$ the Vandesompele et al approach gives (P_{II}, P_I) = (0.28, 0.54) and the direct method gives (P_{II}, P_I) = (0.37, 0.58). When instead $\frac{\sigma_{10}^2}{\sigma_1^2} = 50$ the Vandesomple et al method gives (P_{II}, P_I) = (0.53, 0.39) and the direct method gives (P_{II}, P_I) = (0.54, 0.45). Again we see a clear superiority of the direct method. The above simulation study shows that a sequential approach to identifying control genes does not generally improve the performance of the method.

For the estimation of the intra-group variances by our direct method it is in most situations better to have many genes instead of only a few genes. This can be seen from the variances of the estimates in (1.4). If all the true variances are of equal size, the standard deviation of the estimates is reduced to 80 % when using 5 genes instead of 3 and reduced to 70% when using 10 genes instead of 3. Only in the situations where most variances are small except for one or two very large variances is it better to use a reduced set of genes i.e. to exclude the genes with large variances.

A main assumption for the Vandesompele et al. approach is the independence of the error terms ε_{ij} in (1.1). If two genes are positively correlated then $s^2(i_1,i_2)$ will underestimate $\sigma_{i1}^2+\sigma_{i2}^2$ and the use of M_i will give misleading results. Contrary to this the variance estimates (1.3) are more robust to a lack of independence. Furthermore, when the samples are divided into groups e.g. two groups the approach of Vandesompele et al. can give misleading results if the normalization gene candidates show systematic differences between the two groups. If the variances within the groups are small and all the genes except one, say, show some difference between the two groups, then the optimal candidate with no difference between the two groups will be excluded early on in the Vandesompele et al. approach.

The relationship between the intra- and inter-group variations and the stability value

We use the bladder data to visualize the relationship between the estimated intra- and inter-group variations and the derived “stability value”. Purely for ease of visualization we reduced the complexity of the bladder sample set from three to two groups, the Ta and T2-4 groups and estimated anew the intra- and inter-group variations, and recalculated the “stability value” of all the candidates (Supplementary figure 1). Using only these two groups the inter-group variation corresponds to the difference between the average expression levels of the Ta and the T2-4 group. Furthermore, the estimated intra-group variations can be used to calculate a confidence interval for the difference. Thus, for each candidate the inter-group variation can be depicted as the difference between the Ta and the T2-4 groups, and the intra-group variation can be depicted as a confidence interval for this difference. Thus, by comparing the upper and lower part of Supplementary figure 1 it is evident that the “stability value” consistently reflect the combined effect of the intra- and inter-group variations.

Further lessons can be learned from Supplementary figure 1: first; it makes readily clear that a subset of the candidates is not suited as normalization genes as they show significant inter-group variation e.g. FLOT2 and GAPD, emphasizing the significance of investigating inter-group variation. Second; in the model based approach we assume the average of the inter-group variations to be around zero and to justify this assumption we require the candidates to be selected from a group of genes with no prior expectation of expression difference between groups. Supplementary figure 1 makes it apparent that in the present experiment the assumption is fulfilled as the average of the actual inter-group variations is almost zero (the dashed line).

Reproducibility of the estimated inter- and intra-group variations

We designed a new experiment to evaluate the reproducibility of the estimated inter- and intra-group variations. In this experiment we measured again the expression levels of a number of the candidates CFL1, ACTB, GAPD, and UBC. In parallel we also measured the expression level of four target genes; CD14, FCN1, CCNG2, and NPAS2; known to be differentially expressed between Ta and T2-4 tumors (Dyrskjot et al. 2003). CD14 and FCN1 are down- and CCNG2, and NPAS2 are up-regulated in Ta tumors compared to T2-4 tumors. Primer sequences for these genes can be found in Supplementary table 1. The assay included, in duplicate, a no-template control and a standard curve of four serial dilution points (in steps of 10 fold) of a cDNA mixture, and in triplicates each of the test cDNAs. The sample set consisted of 12 Ta and 14 T2-4 tumors (including the Ta and T2-4 samples of the first sample set). The raw expression values are available as a text file (Supplementary data set 2).

Using the model based approach inter- and intra-group variations of this new data-set were estimated. These data are visualized in Supplementary figure 2. The Figure reveals a prominent similarity between the previous and the newly estimated inter- and intra-group variations demonstrating the reproducibility of the model based strategy. Bearing in mind the new estimations being based on an extended sample set further accentuates the reproducibility.

As expected the target genes distribute two on each side of the candidates verifying their differential expression in Ta and T2-4 tumors. However, Supplementary figure 2 also makes it clear that in an ordinary experiment with one target gene and one normalization gene not all the normalization candidates would provide the correct results. Candidates like HSPCB and TEGT would correctly identify all four targets as differentially expressed while S100A6 and FLOT2 would identify only CD14 and FCN1 as differentially expressed and furthermore would have overestimated the fold change between the Ta and the T2-4 tumors for these two targets.

Supplementary references

Dyrskjot, L., Thykjaer, T., Kruhoffer, M., Jensen, J. L., Marcussen, N., Hamilton-Dutoit, S., Wolf, H., and Orntoft, T. F. Identifying distinct classes of bladder carcinoma using microarrays. Nat Genet, *33:* 90-96, 2003.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol, *3:* RESEARCH0034, 2002.

Supplementary figure 1

Correlation between the estimated inter- and intra-group variations and the derived “stability values”

The upper part of the plot is a visualization of the inter- and intra-group variations of the candidate normalization genes. For ease of visualization the estimated values are based on only the Ta and T2-4 tumors of the original bladder sample set. The circles represent the estimated inter-group variation (expression difference between the Ta and T2-4 tumor groups) and the vertical bars the estimated intra group variation (confidence interval for the difference). A gene showing no inter-group variation has difference between groups value of 0. If the value is > 0 the gene is systematically higher expressed in Ta tumors than in T2-4 tumors and opposite if the value is < 0. The average of the inter-group variations is almost zero (the dashed line).

The lower part of the plot depicts the estimated “stability values” of the candidate normalization genes. Notice the excellent agreement between the size of the “stability values” and the inter- and intra-group variations of the genes.

Supplementary figure 2

Reproducibility of the estimated inter- and intra-group variations

To assess the reproducibility of the estimated inter- and intra-group variations a new experiment was performed on an enlarged set of Ta and T2-4 tumors. Here the expression level of a randomly selected subset of the normalization gene candidates was measured again in parallel with four target genes known to be differentially expressed in Ta and T2-4 tumors. As in Supplementary figure 1 the plot visualizes the estimated variations. Shown in green are the selected candidate normalization

genes, in red the four target genes, and in black the estimations from the first experiment (also shown in Supplementary figure 1).

Supplementary table 1

Primer sequences for four target genes known to be differentially expressed in Bladder Ta and T2-4 tumors

Symbol	Gene name	Accession no.[a]	Locus link	Forward primer	Reverse primer	Amplicon size	Intron spanning
CD14	CD14 antigen	NM_000591	929	CGCTCCGAGATGCATGTG	CCAGCCCAGCGAACGA	62	no
FCN1	Ficolin 1	ENST00000223427	2219	TGCTAGTCTTGTTCCTGCATATCAA	CGGAGAATGGTGAGCTTGTCA	109	yes
CCNG2	cyclin G2	NM_004354	901	GAAGAGAGATTCCAACCTCGAGAA	TCAATCCTGGACACAAAGTGTTATC	85	yes
NPAS2	neuronal PAS domain protein 2	NM_002518.2	4862	TCATCGGATTTTTGCAGAAACA	GAAGGCTTCCAGTCTTGCTGAAT	79	yes

[a] Primer design based on this sequence.