

Analysis of whole-genome microarray replicates using mixed models

Lorenz Wernisch*
School of Crystallography,
Birkbeck College,
London WC1E 7HX, UK

Sharon L. Kendall
Department of Pathology
and Infectious Diseases,
Royal Veterinary College,
London NW1 0TU, UK

Shamit Soneji
School of Crystallography,
Birkbeck College,
London WC1E 7HX, UK

Andreas Wietzorrek
Department of Pathology
and Infectious Diseases,
Royal Veterinary College,
London NW1 0TU, UK

Tanya Parish
Department of
Medical Microbiology,
Barts and the London,
Queen Mary's School of
Medicine and Dentistry,
London E1 2AA, UK

Jason Hinds
Department of
Medical Microbiology,
St Georges Hospital
Medical School,
London SW17 0RE, UK

Philip D. Butcher
Department of
Medical Microbiology,
St Georges Hospital
Medical School,
London SW17 0RE, UK

Neil G. Stoker
Department of Pathology
and Infectious Diseases,
Royal Veterinary College,
London NW1 0TU, UK

*To whom correspondence should be addressed

Abstract

Motivation: Microarray experiments are inherently noisy. Replication is the key to estimating realistic fold-changes despite such noise. In the analysis of the various sources of noise the dependency structure of the replication needs to be taken into account.

Results: We analyzed replicate data sets from a *Mycobacterium tuberculosis trcS* mutant in order to identify differentially expressed genes and suggest new methods for filtering and normalizing raw array data and for imputing missing values. Mixed ANOVA models are applied to quantify the various sources of error. Such analysis also allows us to determine the optimal number of samples and arrays. Significance values for differential expression are obtained by a hierarchical bootstrapping scheme on scaled residuals. Four highly upregulated genes, including *bfrB*, were analyzed further. We observed an artefact, where transcriptional readthrough from these genes led to apparent upregulation of adjacent genes.

Availability: All methods and data discussed are available in the package YASMA <http://www.cryst.bbk.ac.uk/wernisch/yasma.html> for the statistical data analysis system R (<http://www.R-project.org>).

Contact: l.wernisch@bbk.ac.uk

1 Introduction

Microarray technology provides a way to look at gene expression at a whole genome level (Eisen *et al.*, 1998). Analysis of data from microarray experiments is not trivial, and there is a need to identify sources of error in the experimental protocols, in order to provide clear guidance to laboratory scientists so that these new technologies are used effectively.

In this paper we describe the analysis of experiments comparing gene expression in wild-type *Mycobacterium tuberculosis* with that in a defined mutant. A major concern with this slow growing organism is replicability of expression experiments from different cell cultures. mRNA extraction, reverse transcription, hybridization, and scan-

ning are further steps prone to unavoidable variability and noise. Replication is the key to reliable estimation of the amount of differential expression (Lee *et al.*, 2000) and precision in estimation increases with the number of replicates. In practice there is a limit to this number. Different stages of microarray experiments differ considerably in costs; the question arises at which level replication is most effective.

A number of methods have been suggested for the analysis of replicates to extract significance values for differential expression. A widely used approach is based on *t*-statistics for single genes (Dudoit *et al.*, 2000). Significance values derived from *t*-statistics but using a resampling scheme is described in Tusher *et al.* (2001). In contrast, ANOVA based methods estimate variance contributions common to all genes (Kerr & Churchill, 2001; Kerr *et al.*, 2000). It is this approach we take in this paper, although we analyze our experimental design as a mixed model.

2 Results

2.1 Experimental design

We have previously isolated a mutant of *M. tuberculosis* (H37Rv) carrying a defined deletion in the two-component sensor gene *trcS* (Figure 1; submitted for publication). In order to determine transcriptional differences in this mutant, we prepared RNA from cultures of wild-type and mutant bacteria. One culture of each strain (mutant and wild-type) was grown on three separate occasions, resulting in three pairs of cultures, and RNA was extracted from each culture. Fluorescently labeled cDNA was prepared (WT, Cy3; mutant, Cy5) from each pair of cultures, and hybridized to a microarray slide. The cDNA syntheses and hybridizations were repeated, producing a total of six arrays. We performed two spot quantifications on each array to assess variability due to differences in the manual input required by the image analysis software. Thus, three different samples were each prepared and hybridized twice, and each array in turn ana-

lyzed twice resulting in 12 array data sets with signal and background intensities in both channels for 3924 genes.

Our interests are to identify sources of variability and differentially expressed genes. To simplify analysis we mainly consider the logarithm (in base 2) $\log \mathcal{R}/\mathcal{G}$ of the ratio of Cy5 intensity (denoted by the symbol \mathcal{R}) to Cy3 intensity (\mathcal{G}).

2.2 Quality measures for replicated experiments

If fluorescence intensities from different arrays are to be compared, some form of normalization is necessary. We use the following two criteria to assess improvements in consistency between arrays brought about by normalization procedures.

A measure of overall correlation between all 12 experiments is R^2 , the fraction of variance common to the log ratio values of all 12 array sets (see methods). Table 1 contains the R^2 values after application of various filtering and normalization procedures discussed below.

The insertion sequence *IS6110* is present in 16 copies in the *M. tuberculosis* H37Rv genome. This element contains two coding open reading frames (ORFs) (of length 324 bp and 936 bp). In fact these two ORFs represent a single gene, with translational frameshifting causing a single protein to be synthesized. Probes (of length 182 bp and 804 bp) for each ORF are present on the arrays. Transcription of at least one IS element is induced in the mutant, presumably due to a adjacent host promoter that is upregulated. This ORFs constitute a single control present in 32 copies on each array. Effective normalization methods should result in a reduction in the standard deviation of average log ratio values of these copies (see Table 1).

2.3 Data preparation, normalization

The first row in Table 1 shows that there was a low overall correlation and high variance in expression of the insertion sequence when data were not treated (*no bg*). The following four steps were taken to improve data quality.

Background correction. If there is background fluorescence over the whole slide, and hybridization occurs on top of it, spot intensity can be determined by subtracting background from the apparent signal. Problems occur where background levels higher than the spot intensity are seen; here the resulting negative values are corrected to a notional value just above zero. Alternatively, background correction is unnecessary if hybridization takes place efficiently where cDNA finds its complement, excluding background interactions.

As seen in Table 1, overall correlation is higher without (0.57) than with background correction (0.37). However, as the R^2 value after correction for intensity dependence shows (*no bg/lin* or *no bg/loe*, see below), the high R^2 value is an artefact due to correlation in spot intensities in addition to correlation in log ratios. This is also reflected in an increase in variance of *IS6110* genes. We therefore apply background correction.

Removing low intensity points. Since genes with a low spot intensity $1/2(\log \mathcal{R}\mathcal{G})$ result in unreliable log ratio values, we flagged a fraction q of spots with lowest total intensities on each array set. A gene was removed from further consideration if it was flagged on three or more arrays. As seen in Figure 2, the R^2 value peaks at an optimal threshold of $q = 0.2$. In order to retain all 32 insertion genes for the purpose of demonstration, we set the final threshold to $q = 0.07$. 450 genes were discarded at this level and we were left with 3474 genes. Rows with *bg*, *rm* in Table 1 show the improvements.

Imputation of missing data by ANOVA. Values flagged for low intensity (that is, values for genes flagged on only one or two arrays) were imputed so that the overall sum of squared residuals of an ANOVA model was minimized. The particular model used here is detailed in equation 4 of the methods section. We imputed missing values since ANOVA analysis for balanced data is much easier than for unbalanced data.

All in all, 1525 log ratio values for 590 genes were imputed, which is still a small fraction of the total of 83376 log ratio values and is unlikely to distort further statistical results (although the degrees of

freedom in the ANOVA analysis were reduced accordingly). Imputation resulted in a further slight improvement (Table 1, *bg/rm/ip*).

2d normalization. Hybridization is not uniform over the array. It seems to affect dyes differently in different parts of the array in a way that is hard to control experimentally or to correct by general normalization methods. One way to make such systematic differences visible is to fit a trend surface to log ratio values over the arrays. This requires that the position of genes on the array is randomized. We fitted a two-dimensional loess surface to log ratio values on the array sets (two examples are shown in Figure 3). Once a surface had been fitted, log ratio values were corrected by subtracting the surface values. Again an improvement in quality was observed (Table 1, *bg/rm/ip/2d*). This method is related to the pin dependent loess correction suggested by Dudoit *et al.* (2000); but as can be seen in Figure 3, 2d effects are not necessarily confined to pin areas.

Linear and Loess normalization. Alternative approaches to correction such as correction by a local regression curve (Dudoit *et al.*, 2000) (*loe*) or by a linear regression (*lin*) were tested as well. However, there was less improvement in quality (Table 1, *bg/rm/ip/lin*, *bg/rm/ip/loe*).

2.4 Variance and significance analysis

ANOVA is particularly suited for estimating the amount of variation if several experimental steps such as sample growth or mRNA preparation are involved. Effects in an ANOVA model can either be fixed or random. An effect is *fixed* when the set of levels remains the same in future experiments. This is certainly true for the genes, that is, the levels of the gene effect. On the other hand, the 3 samples and 6 arrays are random representatives of a (infinite) population of possible further microarray experiments and are thus *random* effects. A *mixed* model comprises both effects and variability is estimated by *variance components*.

Estimating the amount of variation. The observed variation in log ratio values $\log \mathcal{R}/\mathcal{G}$ for the

same gene in different array sets stems either from sample variation, array variation, or variation in spot quantification. Table 2 shows the variance components σ_X^2 of the corresponding effects X (for an intuitive interpretation of such effects see Figure 4). The G effect (that is, gene averages over arrays) is fixed and has no variance component. The sample effect S (sample averages) is 0 and the array effect A, SA (array averages) is very small; this is due to normalization (note that arrays are nested in samples and the A effect needs to be combined with the interaction SA). Noticeable variation come from the interaction GS (the effect of samples on differential expression of genes), the interaction GA, GSA (the effect of arrays on genes), and the residuals (essentially the effect of spot quantification on genes). The mean square (MS) is *not* a good estimate for the variability of an effect, since it is actually a combination of several variance components as seen in equation 6 of the methods section.

These estimates of variance components provide only global accounts of variation. A closer look at the dependency of variation on average gene intensity reveals an interesting pattern. Figure 5 shows the squares of residuals at each level of replication (as indicated in Figure 4) plotted against average log intensities of genes. Thus the three plots correspond to mean square (MS) values of the the gene-sample effect GS , the gene-array effect GA, GSA , and the residual error. Residual variance is obviously not constant at each level. We therefore split the genes into 3 different regimes of spot intensities and calculated variance components for each regime; they are shown in Table 3.

We would expect sample variation GS for each gene to be independent of spot intensity. This is confirmed in Table 3 and there is only a slight increase for higher intensities. This increase implies that samples vary slightly more for genes that are highly expressed in wild-type and in mutant (we have no biological explanation for this effect, it could be a sampling artefact). On the other hand, as expected, variation of gene-quantification (residuals) and gene-array (GA, GSA) interactions improved dramatically with spot intensity.

Optimal experimental designs. Once variance components have been calculated the following equation provides the variance of the average of differential gene expression over replicates:

$$\text{Var}(\widehat{G}_g) = \frac{1}{n_S} (\sigma_S^2 + \sigma_{GS}^2) + \frac{1}{n_S n_A} (\sigma_{SA}^2 + \sigma_{GSA}^2) + \frac{1}{n_S n_A n_Q} \sigma^2 \quad (1)$$

Here \widehat{G}_g is the average log ratio of gene g , σ_S^2 , σ_{GS}^2 , σ_{GSA}^2 , and σ^2 are the variance components of the corresponding effects (as in Table 2) and n_S , n_A , and n_Q are the number of samples, of arrays per sample, and of quantifications per array (in our case 3, 2, and 2; see Oehlert (2000) for methods of deriving such equations).

If the costs of each component is specified and if an upper limit on the total cost of experiments is imposed, the optimal number of samples S , arrays A , and quantifications Q can be easily derived from equation 1 using the variance components from Table 2. As an example, let us assume the costs for a new wild-type and mutant culture are, say, 50 arbitrary units, for an array 10 units, and for a laser scan and spot quantification 1 unit. It turns out that the optimal design is $S = 3$, $A = 2$, and $Q = 2$; there is no design with the same or smaller variance in estimates for genes that is more cost effective.

Significance of over- and under-expression. True differential expression between mutant and wild-type, to be identified as such, has to rise above the noise level. Variance estimates derived from replicated experiments can be used to obtain a threshold for differential expression. Assuming a normal distribution of residuals, the threshold δ for significant over-expression in averages of log ratios for a gene is

$$\delta = \Phi^{-1} \left(1 - \frac{0.01}{n_G} \right) \sqrt{\text{Var}(\widehat{G}_g)} \quad (2)$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function. Here the overall significance level is 0.01. Since we are testing

$n_G = 3474$ genes, a simple Bonferroni correction for multiple testing reduces the significance level for a single gene to $0.01/n_G$. For our data, this resulted in a δ of about 0.89, that is, a fold-change of about $2^{0.89} = 1.86$ for over- and $1/1.86 = 0.54$ for under-expression.

One of the underlying assumptions in this analysis, constant variance of residuals, is not fulfilled, as seen in Figure 5. Consequently, δ can at best provide an approximate threshold for differential expression. To obtain more reliable estimates, we employ a hierarchical resampling procedure of residuals in which residuals are rescaled according to the fitted curves in Figure 5 (details in the methods section). An alternative strategy would be to use variance stabilizing transformations (Huber *et al.*, 2002).

2.5 Analysis of gene expression in *M. tuberculosis trcS*

Following application of our statistical procedures, we identified a total of 14 over-expressed genes (1.7 to 46-fold, Table 4) and 36 under-expressed genes (0.26 to about 0.6 fold, not shown), with a Bonferroni corrected p -value of 0.01 (IS6110 genes were excluded from this list). Less conservative methods of correction for multiple testing, such as the False Discovery Rate (Benjamini & Hochberg, 1995), result in slightly longer lists (not shown). The most dramatic changes were the up-regulation of four genes. As seen in Figure 6, two of these appear to be an operon, as the stop codon of *mmpS5* (*Rv0677c*) overlaps the predicted start codon of *mmpL5* (*Rv0676c*). A third gene, *Rv0678* is adjacent to these, and divergently transcribed. The fourth gene, *bfrB* (*Rv3841*), lies elsewhere on the genome. Thus there appear to be three highly up-regulated transcriptional units (Figure 6).

The initial aim of these experiments was to identify genes under the control of the transcriptional regulator TrcR. There is evidence that the *trcRS* operon is autoregulated by *trcR* (Haydel *et al.*, 2002). Thus, if the differences we see were directly due to the action of TrcR, we would expect to see a change in *trcR* expression. However,

this was not the case, suggesting that the differences seen have another cause (see discussion). Despite this, the high upregulation of the four genes described above was intriguing.

We partially validated the results using real-time PCR. Primers were designed to the *sigA* gene, which has been used elsewhere to normalize results, and to *bfrB*. mRNA was quantified, and *bfrB* shown to be induced 9.8-fold in the mutant compared to the wild-type strain (not shown). This compares well with the figure of 7.9-fold seen in the arrays.

Transcriptional readthrough leads to expression artefacts. Looking at other over-expressed genes in *M. tuberculosis* *trcS*, it was striking that genes adjacent to the main four (Figure 6) were high on the list *ech5* (*Rv0675*), *glpQ1* (*Rv3842c*) and *Rv0679c* (Table 4), with highly significant up-regulation of over two-fold. However, it is clear from the genome that these all lie at the ends of probable operons, and upstream genes are not up-regulated. The most likely explanation is that although this is real hybridization, it is due to transcriptional run-through from the three main operons.

3 Methods

3.1 Statistical methods

Total correlation between all experiments. As a measure of agreement between experiments we use the following R^2 value. If y_{gk} denotes the log ratio value of gene g on array k , the average for a gene g is an estimate of its true log ratio, $\bar{y}_g = \sum_k y_{gk}/K$ (for K arrays). Agreement between arrays can then be expressed numerically as the amount of variation explained by this average compared to the total variation. If the total mean over all y_{gk} values is μ this fraction is

$$R^2 = \frac{\sum_g K(\bar{y}_g - \mu)^2}{\sum_{g,k} (y_{gk} - \mu)^2} \quad (3)$$

Note that this is the also the R^2 of a one-way ANOVA with genes as treatment groups. We applied R^2 to ranks of log ratio values.

Imputation of missing data. Since balanced designs have closed form solutions, imputing a few missing values is preferable to analyzing the data as nonbalanced design. We suggest imputation of missing values in such a way that the residual sum of squares is minimized. That is, if X denotes the design matrix, \mathbf{y} the data vector (including the missing data as variables), and \mathbf{b} the estimates for effects, then the residual sum of squares $SS_E = (\mathbf{y} - X\mathbf{b})'(\mathbf{y} - X\mathbf{b})$ should be at a minimum. The resulting normal equations

$$\frac{\partial SS_E}{\partial \mathbf{b}} = 2X'\mathbf{y} - 2X'X\mathbf{b} = 0 \quad \text{or} \quad X'\mathbf{y} = X'X\mathbf{b}$$

can be solved for \mathbf{b} . Note that the expression for \mathbf{b} still contains variables for the missing values. In addition, for the variable y_i of the i -th missing data point

$$\frac{\partial SS_E}{\partial y_i} = 2y_i - 2X_i\mathbf{b} = 0 \quad \text{or} \quad y_i = X_i\mathbf{b}$$

with X_i the i -th row of X . Plugging in the above solution for \mathbf{b} results in a system of linear equations in the missing data variables that is solved easily.

For imputation we used the following ANOVA model. After normalization and a logarithmic transformation the values are y_{savgq} for sample s , array a , variety v (mutant or wild-type), gene g , and quantification q . The model is

$$y_{savgq} = \mu + Q_q + S_s + G_g + (S/A)_{sa} + (VG)_{vg} + \epsilon_{savgq} \quad (4)$$

where (S/A) indicates nesting of A in S and ϵ_{savgq} follows a normal distribution.

Analysis of variance components. The sample S and array A , SA (or alternatively, S/A) effect, and consequently GS and GA , GSA (or alternatively, GS/A), are random effects. Let y_{gsaq} be the log ratio value for gene g in sample s on array a and in quantification q . We assumed the following model:

$$y_{gsaq} = \mu + G_g + S_s + (GS)_{gs} + (S/A)_{sa} + (GS/A)_{gsa} + \epsilon_{gsaq} \quad (5)$$

Only the total mean μ and the G effect are fixed, all other effects are random. Random effects are normally distributed with zero mean and variance components σ_S^2 for the S , σ_{GS}^2 for the GS , σ_{SA}^2 for the S/A , σ_{GSA}^2 for the GS/A , and σ^2 for the residual effect. With this notation the expectations of mean squares become

$$\begin{aligned}
 E(MS_S) &= \sigma^2 + n_Q \sigma_{GSA}^2 + n_G n_Q \sigma_{SA}^2 \\
 &\quad + n_{AN} n_Q \sigma_{GS}^2 + n_G n_{AN} n_Q \sigma_S^2 \\
 E(MS_{SA}) &= \sigma^2 + n_Q \sigma_{GSA}^2 + n_G n_Q \sigma_{SA}^2 \\
 E(MS_{GS}) &= \sigma^2 + n_Q \sigma_{GSA}^2 + n_{AN} n_Q \sigma_{GS}^2 \\
 E(MS_{GSA}) &= \sigma^2 + n_Q \sigma_{GSA}^2 \\
 E(MS_E) &= \sigma^2
 \end{aligned} \tag{6}$$

See Oehlert (2000) for methods to derive such equations for a particular design. If mean square values of the ANOVA analysis in Table 2 are used as point estimates of expected mean squares, then equation 6 can be solved for variance components: $\sigma_S^2 = -0.0001$, $\sigma_{G,S}^2 = 0.0606$, $\sigma_{A,S}^2 = 0.0002$, $\sigma_{G,A,S}^2 = 0.0797$, and $\sigma_E^2 = 0.0690$. These are the ANOVA estimates of variance components.

The estimate of σ_S^2 is negative; since it estimates a variance, it should be positive. The value is small and can be safely set to 0. However, in general restricted maximum likelihood estimates (REML) are widely used to obtain non-negative estimates or variance components (Searle *et al.*, 1992). These are maximum likelihood estimates with fixed effects excluded from the maximization process. For balanced factorial designs, REML estimates are identical to ANOVA estimates *if* the latter are non-negative.

In general, REML for linear models involves manipulation of the full covariance matrix of observation variables. This is impractical with models for microarray data that usually contain effects with many thousands of levels. Fortunately, for balanced factorial designs, maximum likelihood estimation can be undertaken on the variance components directly (Thompson, 1962; Thompson & Moore, 1963). We give a brief account of this method.

The mean square M of f independent variables, each normally distributed with mean 0 and variance m , is distributed according to the chi-square distribution

$$\frac{1}{\Gamma(f/2)} \left(\frac{f}{2m} \right)^{f/2} M^{f/2-1} \exp\left(-\frac{fM}{2m}\right) \tag{7}$$

The expectation of M is m . Consequently, this is also the distribution of an ANOVA mean square M_i with f_i degrees of freedom and expected value m_i . Since all mean squares in a balanced factorial design are independent the resulting probability distribution is the product of (7) for all M_i . The log-likelihood for $\mathbf{m} = (m_1, \dots, m_p)$ is

$$l(\mathbf{m}) = \text{const} - \frac{1}{2} \sum_{i=1}^p f_i \left(\log m_i + \frac{M_i}{m_i} \right) \tag{8}$$

under the nonnegativity constraint on variances, $\mathbf{s} = A^{-1} \mathbf{m} \geq 0$, where A is a matrix of coefficients as in equation 6 and M_1, \dots, M_p are the ANOVA mean square values.

Since this function can have several local maxima, straightforward application of a standard optimization procedures results in suboptimal solutions. By imposing constraints obtained from the Kuhn-Tucker theorem, an optimal numerical solution can nevertheless be easily obtained. For a local maximum of function $l(\mathbf{m})$

$$A^{-1} \mathbf{m} \geq 0, \quad A' \nabla l(\mathbf{m}) \leq 0, \quad \mathbf{m} \nabla l(\mathbf{m}) = 0 \tag{9}$$

where $\nabla l(\mathbf{m})$ is the gradient of l at \mathbf{m} . If \mathbf{b}_i denotes the rows of matrix A^{-1} and \mathbf{a}_i the rows of matrix A' this is equivalent to

$$\begin{aligned}
 \text{for each } i, \text{ either } & \mathbf{b}_i \mathbf{m} = 0, \mathbf{a}_i \nabla l(\mathbf{m}) \leq 0 \\
 \text{or } & \mathbf{b}_i \mathbf{m} > 0, \mathbf{a}_i \nabla l(\mathbf{m}) = 0
 \end{aligned} \tag{10}$$

If components can be identified that are zero, either from negative ANOVA estimates or in the first minimization round, very precise variance estimates can be obtained by enforcing $\mathbf{b}_i \mathbf{m} = 0$ for these rows i and $\mathbf{a}_j \nabla l(\mathbf{m}) = 0$ for all other rows j . The result is shown in Table 2; σ_S^2 actually is $1.1 \cdot 10^{-16}$ and σ_{AS}^2 reduced from an estimate by ANOVA of 0.00020 to one by REML of 0.00013.

Bootstrap sampling of residuals. For the significance analysis we generated random data sets by resampling from the original data and use them for p -value estimates. Residuals at the various levels, as shown in figures 4 and 5, were randomly shuffled between genes, but only within each level and with an appropriate scaling derived from the curves in Figure 5. Repeated application of the shuffling procedure generated 500 new data sets. For each data set we calculated log ratio averages \widehat{G}_g^* for each gene g . The values \widehat{G}_g^* fluctuate around the original log ratio averages \widehat{G}_g . The bootstrap values are close to a normal distribution (a typical p -value in the Shaprio-Wilk test is 0.57 for *Rv1884c*, for example). Hence, p -values for differential expression were obtained from a fitted Gaussian distribution. The result for over-expressed genes are shown in Table 4.

3.2 Laboratory protocols

Microarray construction. Whole genome microarrays were constructed by robotic spotting onto poly-lysine coated glass microscope slides (Micro-Grid II, BioRobotics, UK) of PCR amplicons (size range 60 - 1000 bp; mean: 517bp) derived from portions of each of the 3924 predicted ORF's of the sequenced strain of *M. tuberculosis* H37RV. Primer pairs for each ORF were designed with Primer 3 software and selected by BLAST analysis to have minimal cross-homology with all other ORF's. All procedures used including post-processing of deposited arrays and hybridization of slides were as described by others (Wilson *et al.*, 2001).

Preparation and labeling of cDNA. For RNA isolation, 100ml cultures in were grown with constant rolling in Middlebrook 7H9 liquid medium supplemented with 10% (v/v) OADC (Becton Dickinson) plus 0.05% Tween 80 to mid-log (7 days). RNA was isolated using a commercially available kit (Qiagen), and treated twice with DNase. 10 μ g of total RNA from wild-type and mutant bacteria were used in independent labeling reactions. Mutant RNA was labeled with Cy5-dCTP and wild-type RNA with Cy3-dCTP.

Scanning. Slides were scanned with a Genetic Mi-

croarrays GMS 418 array scanner following the manufacturer's guidelines. Fluorescent spot intensities were quantified using ImaGene 4.1 (BioDiscovery Inc.) software.

4 Discussion

In this paper we present the analysis of microarray experiments comparing an *M. tuberculosis* mutant with the wild-type. One concern is the identification and estimation of the different sources of variability in replicated microarray experiments. A second concern is providing significance thresholds for differential gene expression. We suggest new methods to answer both of these issues.

For the data analyzed here the most effective approach for data preparation and normalization was to remove spot background, remove low intensity genes, impute missing data, and apply 2d normalization. Low intensity values for genes not removed were imputed via an ANOVA model. Other methods for imputing missing microarray data have been suggested by Troyanskaya *et al.* (2001), but they are less suitable for our hierarchical replications.

Mixed model analysis of microarray data has been described, for example, in Wolfinger *et al.* (2001). Our model differs from Wolfinger's in that we introduce common variance components for all genes and consider the effects of a hierarchical replication design. Equation 1 provides the basis for calculating the number of samples, arrays, or other replication units needed to lower variability in a cost-effective way. Whereas Lee *et al.* (2000) suggest a fixed experimental design, such numbers depend on observed variability in the arrays and the costs of work and material. In order to focus on the main sources of variability, we did not consider other possible effects, such as dye-gene effects, in the experimental design.

The most widely used method for significance analysis of differential expression is the application of t -tests to replicated values of individual genes (for example, Dudoit *et al.* (2000)). In contrast, ANOVA analysis as presented here (see also Kerr *et al.* (2000)), derives combined error estimates

for all genes simultaneously. The advantage of the ANOVA approach is that the degrees of freedom used in the variance estimates are large and estimates become more reliable. The price to pay is lack of specificity for single genes.

The underlying assumptions in an ANOVA analysis of normality and constant variance are problematic as well. Various statistical models have been suggested that go beyond this assumption (Newton *et al.*, 2001; Baggerly *et al.*, 2001; Huber *et al.*, 2002). Expression data might also be modelled without any parametric assumptions (Efron *et al.*, 2000; Tsodikov *et al.*, 2002). Nevertheless, for the data presented here simplifying assumptions seem to be justified, in particular after proper normalization. Their advantage is that they allow us to focus on the effects of the hierarchical experimental scheme. For a significance analysis we do not rely on the variance being constant, as we employ a bootstrapping scheme of residuals that takes dependence of variance on intensity into account. This method combines the best of both worlds, estimation of variances based on large samples and sensitivity to characteristics of individual genes such as their spot intensity.

In these experiments we used spotted arrays. The factors analyzed in our model of variance components (sample, array, image analysis) will also apply to Affymetrix data. Normalization issues though are completely different.

One potential phenomenon revealed by our analysis was that transcription from highly expressed genes may extend at least partly into the neighbouring ORF, without leading to expression of that gene. Hybridization to the PCR product of such gene may lead to erroneous conclusions about its expression. The use of arrayed oligonucleotides may avoid this problem.

References

- Baggerly, K. A., Coombes, K. R., Hess, K. R., Stivers, D. N., Abruzzo, L. V. & Zhang, W. (2001) Identifying differentially expressed genes in cDNA microarray experiments. *J. Comput. Biol.*, **8**, 639–659.
- Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B*, **75** (1), 289–300.
- Dudoit, S., Yang, Y. H., Callow, M. J. & Speed, T. P. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578 University of California at Berkeley. <http://stat-ftp.berkeley.edu/tech-reports> .
- Efron, B., Tibshirani, R., Goss, V. & Chu, G. (2000). Microarrays and their use in a comparative experiment. Technical report Stanford University. <http://www-stat.stanford.edu/~tibs/research.html> .
- Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 14863–14868.
- Haydel, S. E., Benjamin, Jr, W. H., Dunlap, N. E. & Clark-Curtiss, J. E. (2002) Expression, autoregulation, and DNA binding properties of the Mycobacterium tuberculosis TrcR response regulator. *J. Bacteriol.*, **184**, 2192–2203.
- Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. & Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. In *Proceedings of the 10th Int. Conference on Intelligent Systems for Molecular Biology (ISMB02)* Oxford University Press.
- Kerr, M. K. & Churchill, G. A. (2001) Statistical design and the analysis of gene expression microarray data. *Genet. Res.*, **77**, 123–128.
- Kerr, M. K., Martin, M. & Churchill, G. A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

- Lee, M. L., Kuo, F. C., Whitmore, G. A. & Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 9834–9839.
- Newton, M. A., Kendzierski, C. M., Richmond, C. S., Blattner, F. R. & Tsui, K. W. (2001) On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *J. Comput. Biol.*, **8**, 37–52.
- Oehlert, G. W. (2000) *A First Course in Design and Analysis of Experiments*. W. H. Freeman, New York.
- Pinheiro, J. C. & Bates, D. M. (1996) Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, **6**, 289–296.
- Searle, S. R., Casella, G. & McCulloch, C. E. (1992) *Variance Components*. John Wiley.
- Thompson, W. A. (1962) The problem of negative estimates of variance components. *Annals of math. statistics*, **33**, 273–289.
- Thompson, W. A. & Moore, J. R. (1963) Non-negative estimates of variance components. *Technometrics*, **5**, 441–449.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R. B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics.*, **17**, 520–525.
- Tsodikov, A., Szabo, A. & Jones, D. (2002) Adjustments and measures of differential expression for microarray data. *Bioinformatics.*, **18**, 251–260.
- Tusher, V. G., Tibshirani, R. & Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 5116–5121.
- Wilson, M., Voskuil, M., Schnappinger, D. & Schoolnik, G. K. (2001) Functional genomics of *Mycobacterium tuberculosis* using DNA microarrays. In *Methods in Molecular Medicine, vol 54: Mycobacterium tuberculosis Protocols* pp. 335–357 Humana Press Inc, Totowa, NJ.
- Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. & Paules, R. S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, **8**, 625–637.

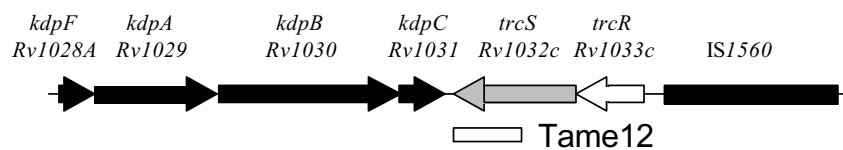


Figure 1: The regulator-sensor pair *trcS* (*Rv1032c*) and *trcR* (*Rv1033c*). The open bar shows the DNA deleted in the mutant Tame12.

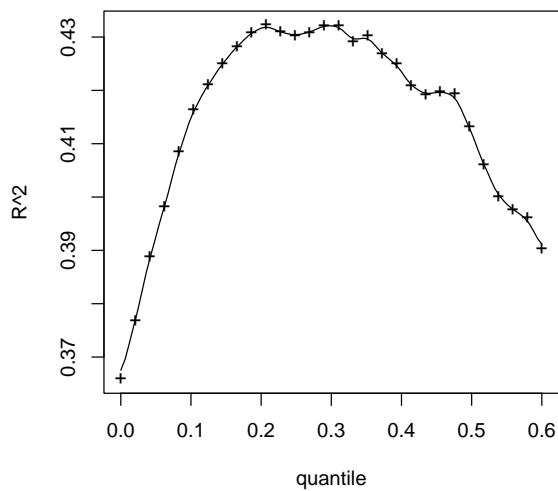


Figure 2: Overall rank correlation R^2 when a fraction of low intensity spots is flagged on each array set and genes with at least three flagged spots are removed.

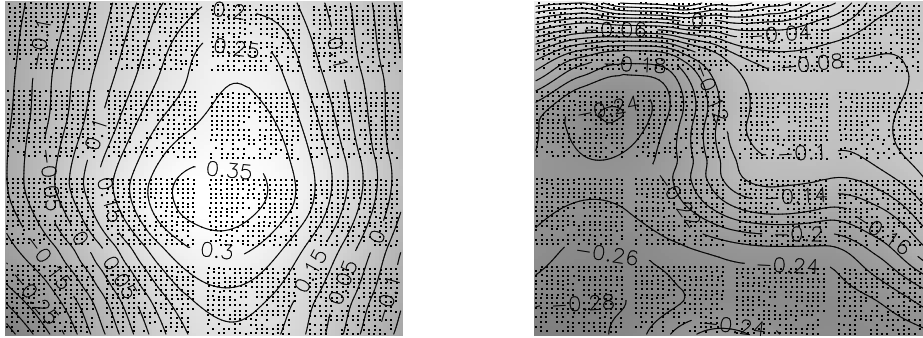


Figure 3: Two-dimensional loess surface (degree 1, span 0.1) fitted to log ratio values from array set 10 and from array set 12. Over-expression is indicated by brighter shades. The location of spots on the array is indicated by dots.

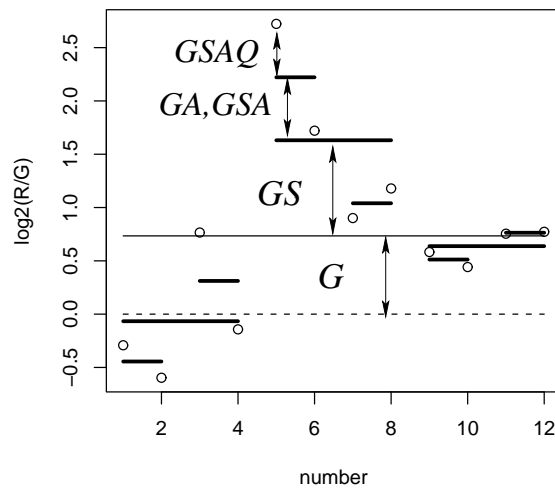


Figure 4: Various ANOVA effects of a single gene. Compare with Table 2.

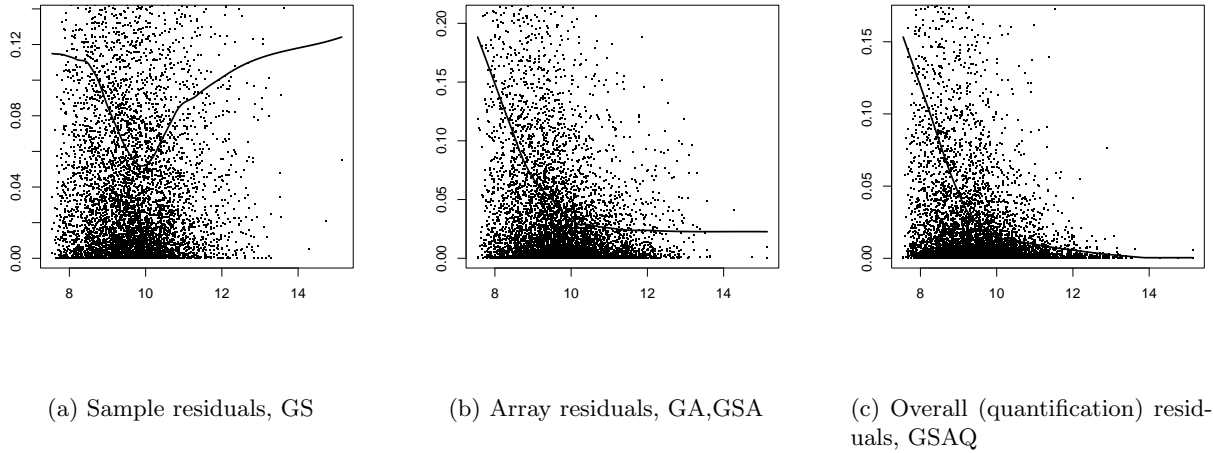


Figure 5: Squares of residuals over average log intensities $1/2 \log \mathcal{RG}$ of genes. The plots show residuals of samples over genes, of arrays over samples, and of quantifications over arrays. Compare with Figure 4. Loess fit is with degree 1 and span 0.4.

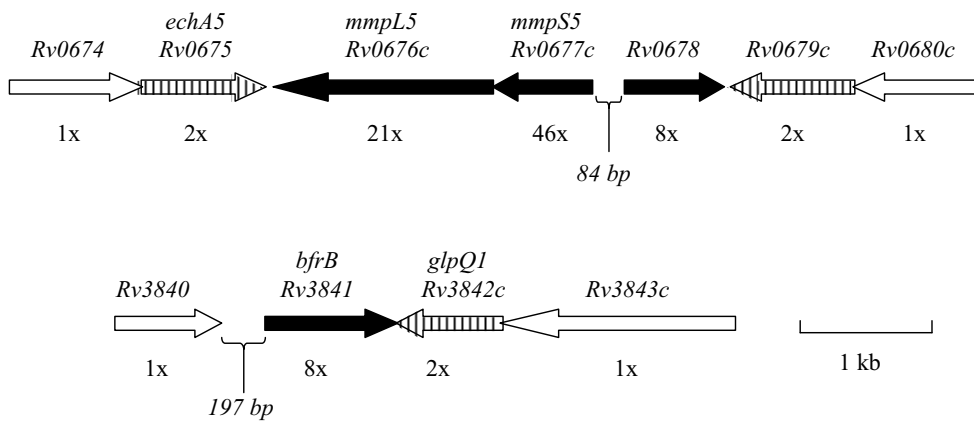


Figure 6: The gene context of the four top up-regulated genes *mmpL5*, *mmpS5*, *Rv0678*, and *bfrB*.

Method	overall R^2	sd of IS6110	range in fold change IS6110
no bg	0.57	0.41	0.63–1.40
no bg, lin	0.34	0.15	1.09–1.66
no bg, loe	0.33	0.13	1.09–1.61
bg	0.37	0.21	1.09–2.34
bg, rm	0.40	0.21	1.09–2.34
bg, rm, ip	0.42	0.17	1.41–2.34
bg, rm, ip, 2d	0.43	0.11	1.68–2.36
bg, rm, ip, lin	0.39	0.19	1.51–2.62
bg, rm, ip, loe	0.38	0.23	1.37–2.5

Table 1: Comparison of normalization procedures.

effects	df	MS	σ_X^2
gene G	3473	1.726	
sample S	2	0.3346	0
gene-sample GS	6946	0.4711	0.0606
array A, SA	3	1.613	0.0001
gene-array GA, GSA	10419	0.2285	0.0797
residuals	20844	0.0691	0.0691

Table 2: ANOVA table with variance components.

effects	σ_X^2 low	σ_X^2 middle	σ_X^2 high	σ_X^2 total
sample S	0.011	0.001	0.017	0
gene-sample GS	0.042	0.041	0.070	0.061
array A, SA	0.006	0.001	0.006	0.000
gene-array GA, GSA	0.121	0.069	0.036	0.080
residuals	0.141	0.046	0.020	0.069
$\text{Var}(\hat{G})$	0.051	0.038	0.030	0.039
fold-change	1.95/0.51	1.78/0.56	1.66/0.60	1.86/0.54

Table 3: Variance components for different spot intensity ranges.

	gene	p -value	fold-change
1	<i>mmpS5</i> (<i>Rv0677c</i>)	$1.34e - 172$	46.31
2	<i>mmpL5</i> (<i>Rv0676c</i>)	$5.06e - 142$	21.41
3	<i>Rv0678</i>	$2.16e - 70$	8.46
4	<i>bfrB</i> (<i>Rv3841</i>)	$1.05e - 51$	7.92
5	<i>Rv3130c</i>	$2.11e - 15$	2.63
6	<i>echA5</i> (<i>Rv0675</i>)	$1.03e - 08$	2.20
7	<i>glpQ1</i> (<i>Rv3842c</i>)	$2.34e - 08$	2.27
8	<i>Rv3407</i>	$1.29e - 06$	2.10
9	<i>Rv1398c</i>	$2.26e - 05$	2.12
10	<i>Rv0679c</i>	$4.91e - 05$	2.24
11	<i>Rv1884c</i>	$3.34e - 04$	1.74
12	<i>hspX</i> (<i>Rv2031c</i>)	$3.76e - 03$	1.94
13	<i>PE.PGRS</i> (<i>Rv0109</i>)	$6.22e - 03$	1.66
14	<i>Rv3768</i>	$7.32e - 03$	1.74
15	<i>Rv2147c</i>	$7.49e - 03$	1.79
16	<i>Rv1109c</i>	$9.56e - 03$	1.80

Table 4: Over-expressed genes.