Gene profiling for determining pluripotent genes in a time course microarray experiment

J. TUKE∗, G. F. V. GLONEK, P. J. SOLOMON

School of Mathematical Sciences,
The University of Adelaide, Adelaide SA 5005, Australia
simon.tuke@adelaide.edu.au

SUMMARY

In microarray experiments, it is often of interest to identify genes which have a prespecified gene expression profile with respect to time. Methods available in the literature are, however, typically not stringent enough in identifying such genes, particularly when the profile requires equivalence of gene expression levels at certain time points. In this paper, the authors introduce a new methodology, called gene profiling, that uses simultaneous differential and equivalent gene expression level testing to rank genes according to a prespecified gene expression profile. Gene profiling treats the vector of true gene expression levels as a linear combination of appropriate vectors, for example, vectors that give the required criteria for the profile. This gene profile model is fitted to the data, and the resulting parameter estimates are summarized in a single test statistic that is then used to rank the genes. The theoretical underpinnings of gene profiling (equivalence testing, intersection–union tests) are discussed in this paper, and the gene profiling methodology is applied to our motivating stem-cell experiment.

Keywords: Gene expression; Gene profiling; Linear model; Microarray; Pluripotency; Stem cell; Time course experiment.

1. INTRODUCTION

Microarray technology enables researchers to examine the expression levels for many thousands of genes simultaneously (see, e.g. Nguyen and others, 2002; Smyth and others, 2003). Increasingly, information on gene expression is used to infer cell protein levels and thus cellular behavior (Nguyen and others, 2002; Smyth and others, 2003; Ahnert and others, 2006; McLachlan and others, 2006). A further major area of interest is in investigating changes in gene expression levels over time in a population of cells (Dudoit and others, 2002; Bar-Joseph and others, 2003; Glonek and Solomon, 2004; Tai and Speed, 2006; Ernst and others, 2005; Brown and others, 2006; Ahnert and others, 2006), and this is the subject of the present paper. We refer to the gene expression levels over time as a gene expression profile or profile for short.

∗To whom correspondence should be addressed.

© The Author 2008. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.
Gene profiling for determining pluripotent genes

Several methods of analyzing gene expression profiles fall into the class of techniques known as unsupervised learning methods. These methods seek to group genes into a number of classes based upon their observed profiles. Some of the methodologies discussed in the recent microarray literature are hierarchical classification (Eisen and others, 1998), self-organizing maps (Tamayo and others, 1999), K-means algorithm (Tavazoie and others, 1999), multivariate Gaussian mixtures (Ghosh and Chinnaiyan, 2002; Yeung and others, 2001), and mixtures of linear mixed models (Celeux and others, 2005). A related problem that arises in applications of microarray time course experiments is to specify, in advance, a gene expression profile of interest and then to identify the genes with matching expression profiles. However, unsupervised methods do not address this problem, and various alternative approaches have been proposed.

One such method is Pareto optimization, proposed by Fleury and others (2002) and Hero and Fleury (2004), in which a set of functions, each measuring the association of a gene to a prespecified profile, is chosen. Genes found to be Pareto optimal with respect to these criteria are identified as matching the prespecified profile. The main disadvantage of Pareto optimization is that some genes will be selected as Pareto-optimal genes while only matching the prespecified profile for a subset of the profile’s criteria.

Lönnstedt and others (2001) describe a different method for ranking genes, based on the inner product between the vector of observed log ratios and a prespecified profile. They wanted to isolate genes that had a late response to the treatments applied: oncostatin M (OSM) alone, epidermal growth factor (EGF) alone, or both OSM and EGF. They used a function of time, $f(t) = t^2$, for the desired trend of genes that were late responders. The time measurements in the experiment were 0.5, 1, 4, and 24 h. Therefore, the vector of values, $u$, of genes matching the trend was (0.25, 1, 16, 576). If the vector of a gene’s estimated expression levels for a particular treatment is represented by $y$, then a large dot product $u \cdot y$ would indicate a gene following the prespecified trend. Note that the method does not standardize for within-gene variability. This method works well for some profiles but did not provide useful outcomes in our application.

Gene profiling aims to identify genes that match a prespecified gene expression profile with greater specificity than the previously described approaches. Gene profiling treats the vector of true gene expression levels for each gene as a linear combination of linearly independent vectors chosen to represent the prespecified profile. The gene profile model is fitted to the observed log ratios, and the genes are ranked by a single test statistic that incorporates simultaneous differential and equivalent gene expression testing.

Section 2 describes the experimental design for a pluripotent (stem-cell) time course experiment which provided our initial motivation. The theoretical underpinnings of gene profiling are described in Section 3. The results obtained from our application to a stem-cell experiment are presented in Section 4, where we also compare gene profiling with Pareto optimization and the Lönnstedt method. In Section 5, we discuss further work and show how to apply the methods in limma.

2. Motivation: pluripotency

Our motivating example is a stem-cell experiment conducted by the Rathjen Laboratory, formerly of the University of Adelaide. The aim of the experiment was to identify genes associated with pluripotency in mice embryonic stem cells (D’Amour and Gage, 2003; Ramalho-Santos and others, 2002). Early stem cells have the potential to differentiate into any body cell: a property known as pluripotency. This ability is present in mice stem cells up to and including day 3. After this the stem cells become multipotent: they still have the ability to differentiate into different types of cells but now a limited number. For example, haemopoietic stem cells can differentiate into blood cells but not nerve cells. As pluripotency is restricted to the early stem cells, day 3 or earlier, genes that have high expression levels in cells up to day 3, but low or monotonically decreasing expression levels thereafter, are likely to be associated with the biochemical pathways involved in the pluripotency ability of these cells (C. Wilkinson, personal communication).
2.1 Pluripotency example: experimental design

Stem cells were isolated from the early embryo and grown in culture dishes. The cells were allowed to replicate and grow over the medium in the dish. Once the cells had crowded the plate, they were removed, separated, and plated onto new plates. This cycle of growth and replating is called a “passage.” The Rathjen Laboratory isolated mice embryonic stem cells and for this experiment used cells from passages 21, 22, 23, and 24. The cells were stimulated to differentiate into multipotent cells, and on days 0, 3, 6, and 9 after stimulation samples were taken and the messenger RNA (mRNA) was obtained. The gene expressions were measured for the resulting 16 samples of stem-cell mRNA. Within each passage, 5 comparisons were made, namely, day 0 to day 3, day 0 to day 9, day 3 to day 6, day 3 to day 9, and day 6 to day 9. Optimal experimental design considerations indicate that the most efficient time course experiment from a statistical viewpoint would be one which includes all pairwise comparisons of time points especially for an “omnibus” experiment of this nature intended to address several biological hypotheses of interest in a single experiment (Yang and Speed, 2003; Glonek and Solomon, 2004). However, the experiment was limited to a total of 20 hybridizations, and the individual researchers involved in the collaboration were concerned to ensure that they had sufficient replication of the particular time points of interest for their separate research hypotheses, which for the most part did not include what was happening at day 6 of the development. The design used is not “optimal” for the pluripotent profile discussed in the present paper but is close to optimal (Glonek and Solomon, 2004).

The clone library used in the experiment was the Compugen 22 000 mouse oligonucleotide library (http://www.microarray.adelaide.edu.au/libraries/). In total, 20 arrays were hybridized on 2-color long oligonucleotide microarrays, with 5 arrays within each passage, 1 for each of the 5 comparisons detailed above. In this analysis, the stem cells from each passage were treated as independent biological replicates.

3. Gene profiling methodology

3.1 Development of method for stem-cell experiment

The expression criteria over time required for a pluripotent gene are

• equal gene expression levels for days 0 and 3,
• higher gene expression levels for days 0 and 3 compared to day 9, and
• the gene expression level for day 6 to lie between the gene expression levels for day 0 and day 3, and the gene expression level for day 9.

An example of the requisite profile is illustrated in Figure 1. Consider the vector of true mean gene expression levels \( \mu = (\mu_0, \mu_3, \mu_6, \mu_9)' \), where \( \mu_i \), \( i = 0, 3, 6, \text{ and } 9 \), is the mean gene expression level on day \( i \) as shown in Figure 1. Since this is a vector in \( \mathbb{R}^4 \), it can be expressed as the linear combination of 4 linearly independent vectors. The first step in gene profiling is to choose vectors that represent the criteria for pluripotency. In the present example, this corresponds to

\[
\mu = \gamma_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \gamma_1 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \gamma_2 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} + \gamma_3 \begin{pmatrix} 1/2 \\ 1/2 \\ 0 \\ 0 \end{pmatrix}.
\]

(3.1)

With this choice of model, it follows that \( \gamma_0 = \mu_9 \), \( \gamma_1 = (\mu_0 + \mu_3)/2 - \mu_6 \), \( \gamma_2 = \mu_6 - \mu_9 \), and \( \gamma_3 = \mu_0 - \mu_3 \). Therefore, the pluripotent profile requires that \( \gamma_1 > 0, \gamma_2 > 0 \), and \( \gamma_3 = 0 \) but does not constrain \( \gamma_0 \). To find genes that achieve these criteria requires tests for equivalence as well as (simultaneous) tests...
Fig. 1. The prespecified gene expression profile for pluripotent genes. For each day, the log ratio with respect to day 0 is plotted.

for differential gene expression. In Section 3.2, equivalence testing is discussed. We then describe how to simultaneously test for both differential and equivalent gene expression in a time course experiment.

3.2 Statistical equivalence

To determine pluripotency, it is necessary to demonstrate that $\gamma_3 = 0$. Conventional hypothesis testing is not applicable to this situation, but we can use the equivalence testing approach discussed in Wellek (2002).

If $X$ is a random vector whose probability distribution depends on a real-valued parameter $\theta$, then to test if $\theta$ is equivalent to zero, a neighborhood around zero is constructed, and the following null and alternative hypotheses are tested:

$$H_0: |\theta| \geq \epsilon, \quad \epsilon > 0,$$

$$H_a: |\theta| < \epsilon.$$  

The neighborhood defined by $\epsilon$ is the maximum that the parameter can vary and still be considered equivalent to zero. This neighborhood is necessary to ensure that the power of the statistical test is greater than its significance level (Wellek, 2002).

For the gene profiling model, the parameter $\epsilon$ is taken to be the largest that a gene’s mean log ratio can vary around zero and not be biologically significant. Ideally, a working understanding of equivalent gene expression should be decided upon in advance in consultation with biologists, and this information is used to decide on an appropriate value of $\epsilon$. We discuss choices of $\epsilon$ in Section 4, but for the present we will assume an appropriate $\epsilon$ to be available. Using such a value of $\epsilon$, the simplest and most common way to test the hypotheses in (3.2) is via “confidence interval inclusion” (CII).

Consider the null and alternative hypotheses specified in (3.2). We calculate a confidence interval, $R_{\alpha}(X)$, from the observed data $X$, where

$$R_{\alpha}(X) = (L_{\alpha}(X), U_{\alpha}(X)),$$  

(3.3)
$L_\alpha(X)$ and $U_\alpha(X)$ are random variables such that

$$P(\theta \in (L_\alpha(X), \infty)) = P(\theta \in (-\infty, U_\alpha(X))) = 1 - \alpha.$$  

We reject the null hypothesis in favor of equivalence if and only if

$$R_\alpha(X) \subset (-\epsilon, \epsilon),$$

that is, the confidence interval is contained entirely within the interval $( -\epsilon, \epsilon )$. This is an $\alpha$-level test.

The equivalence formulation can be used to test that $\gamma_3$ in (3.1) is equivalent to zero with the following null and alternative hypotheses:

$$H_0: |\gamma_3| \geq \epsilon \quad \text{versus} \quad H_a: |\gamma_3| < \epsilon.$$  \hspace{1cm} (3.4)

For example, to test the hypotheses in (3.4), the confidence interval $(\hat{\gamma}_3 - t^* \text{SE}(\hat{\gamma}_3), \hat{\gamma}_3 + t^* \text{SE}(\hat{\gamma}_3))$ is calculated, and $\gamma_3$ is concluded to be equivalent to zero if this confidence interval lies within $( -\epsilon, \epsilon )$, with $t^*$ chosen such that $P(T > t^*) = \alpha$, where $T$ has a $t$-distribution with the appropriate degrees of freedom for $\gamma_3$.

CII can also be used to test separately whether $\gamma_1$ and $\gamma_2$ are significantly positive. The null and alternative composite hypotheses for $\gamma_1$ are

$$H_0: \gamma_1 \leq 0 \quad \text{versus} \quad H_a: \gamma_1 > 0,$$  \hspace{1cm} (3.5)

and similarly for $\gamma_2$.

For an $\alpha$-level test here, a one-sided $(1 - \alpha)100\%$ confidence interval for $\gamma_1$ is calculated as $(\hat{\gamma}_1 - t^* \text{SE}(\hat{\gamma}_1), \infty)$ and $H_0$ rejected if this interval is contained in $(0, \infty)$.

These methods allow testing of each criterion separately, but for pluripotency all 3 criteria need to be valid simultaneously. Our method for doing so is described in Section 3.3.

### 3.3 Intersection–union test

The tests for each criterion discussed in Section 3.2 can be incorporated into a single null and a single alternative hypothesis as follows:

$$H_0: (\gamma_1 \leq 0) \bigcup (\gamma_2 \leq 0) \bigcup (|\gamma_3| \geq \epsilon), \quad \epsilon > 0,$$  \hspace{1cm} (3.6)

versus

$$H_a: (\gamma_1 > 0) \bigcap (\gamma_2 > 0) \bigcap (|\gamma_3| < \epsilon).$$  \hspace{1cm} (3.7)

The hypotheses in (3.6) and (3.7) represent an “intersection–union test” (IUT) (Berger, 1982). To review, in an IUT, the null hypothesis is expressed as a union

$$H_0: \theta \in \bigcup_{\gamma \in \Gamma} \Theta_\gamma,$$

where $\Theta_\gamma$ is a subset of the parameter space indexed by $\gamma$. The rejection region is of the form $R = \bigcap_{\gamma \in \Gamma} R_\gamma$, where $R_\gamma$ is the rejection region for a test of $H_{0\gamma}: \theta \in \Theta_\gamma$ versus $H_{1\gamma}: \theta \in \Theta_\gamma^c$. This is an $\alpha$-level test, where $\alpha = \sup_{\gamma \in \Gamma} \alpha_\gamma$, and $\alpha_\gamma$ is the size of the test $H_{0\gamma}$, with rejection region $R_\gamma$.

Thus, for each $\gamma_i, i = 1, 2, 3$, in the null hypothesis statement (3.6), a test of size $\alpha_i$ is found, and the overall IUT will be of level $\sup \alpha_i$. Using the CII method discussed in Section 3.2 to test each $\gamma_i$ separately therefore gives an overall $\alpha$-level test.
Our main aim was to rank the genes in our motivating example according to their match with the pluripotent profile. The testing methodology described can be modified to give a quantitative measure of how closely each gene matches the desired profile. Considering each gene separately, for each parameter \( \gamma_i \), \( i = 1, 2, 3 \), CII is used to test the associated null hypothesis. Rather than using a fixed significance level, for each \( \gamma_i \), \( i = 1, 2, 3 \), the smallest significance level, \( \alpha_i \), is found such that the null hypothesis is rejected. The supremum of \( \alpha_i \), \( i = 1, 2, 3 \), is used as the test statistic to rank the genes. In fact, in the stem-cell experiment, rather than calculating \( \alpha_i \) for each \( \gamma_i \), \( i = 1, 2, 3 \), the width of the largest confidence interval, \( U_i \), for each \( \gamma_i \) that was contained within the rejection region was used. The infimum, \( U \), of the \( U_i \) was then used to rank the genes. This is equivalent to ranking based on \( \sup \alpha_i \).

3.4 Gene profiling for pluripotency

The scanned images for each hybridized microarray slide were analyzed using SPOT (Yang and others, 2001) to give the cy3 and cy5 intensities for each gene (Yang and others, 2001; Adams and Bischof, 1994). The data were then normalized by within-array print-tip loess, and the gene-profile model was fitted to the normalized data using limma (Smyth, 2005) in R (R Development Core Team, 2006). The normalized data are available from the supplementary material available at Biostatistics online (http://www.biostatistics.oxfordjournals.org). For each gene, the model parameter estimates and standard errors obtained by limma were used to calculate the \( U \) statistic (see below) using C code embedded in R code (Section A and B of the supplementary material available at Biostatistics online http://www.biostatistics.oxfordjournals.org). The genes were then ranked using the \( U \) statistic.

The vector of observed log ratios \( M \) was expressed as a linear model of the true gene expression levels \( \mu \) as follows:

\[
M = X^* \mu + E,
\]

where \( X^* \) is the design matrix representing the mRNA comparisons made on each array and \( E \) is assumed to be distributed as \( N_{20}(0, \sigma^2 I) \). Using (3.1) to substitute for \( \mu \) gives

\[
M = X^* \left( \begin{array}{ccc}
1 & 1 & 1/2 \\
1 & 1 & -1/2 \\
1 & 0 & 0 \\
1 & 0 & 0
\end{array} \right) \left( \begin{array}{c}
\gamma_0 \\
\gamma_1 \\
\gamma_2 \\
\gamma_3
\end{array} \right) + E
\]

\[
= X^* \gamma + E.
\]

In the stem-cell experiment, the microarray platform used was 2-color long oligonucleotide which, as for complementary DNA microarrays, measures relative rather than absolute gene expression. Therefore, the overall gene expression level, \( \gamma_0 \), could not be estimated and was removed from the model by changing the parameter vector to \( (\gamma_1, \gamma_2, \gamma_3) \) and removing the first column of \( X \).

Estimates of \( \gamma \) were calculated by least squares, and the estimate of \( \sigma^2 \) was obtained using the empirical Bayes method utilized in limma; this gives a robust posterior estimate of \( \sigma^2 \) based on a prior which “borrows” information from the observed variance of all the genes on the array.

For each gene, 3 test statistics, \( U_1, U_2, \) and \( U_3 \), were calculated as follows:

\[
U_1 = \frac{\hat{\gamma}_1}{\text{SE}(\hat{\gamma}_1)}, \quad U_2 = \frac{\hat{\gamma}_2}{\text{SE}(\hat{\gamma}_2)}, \quad U_3 = \frac{\epsilon - |\hat{\gamma}_3|}{\text{SE}(\hat{\gamma}_3)},
\]

where \( \text{SE}(\hat{\gamma}_i) \) is the \( i \)th diagonal element of the square matrix \( s \sqrt{(X'X)^{-1}} \), and \( s \) is the posterior estimate of \( \sigma \). The minimum of \( U_i, i = 1, 2, 3 \), is used to rank the genes. Genes whose estimate \( (\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3) \) of
did not lie within the rejection region, that is, those genes for which at least one $U_i, i = 1, 2, 3,$ was negative, were excluded from the ranking.

4. APPLICATION: DETERMINING GENES ASSOCIATED WITH PLURIPOTENCY USING GENE PROFILING

The model (3.1) was fitted to the stem-cell data with $\epsilon = 1$. In addition, the test statistics were changed to test for $\gamma_2 > 1.5$, that is, $U_2 = (\hat{\gamma}_2 - 1.5)/SE(\hat{\gamma}_2)$. The value of 1.5 was chosen to ensure a large difference between the gene expression levels on days 0, 3, and 6 compared with the gene expression level on day 9.

The ranked genes are given in Table 1, and the fitted profiles for these 15 genes are shown in Figure 2.

Table 1. The ranked genes from fitting pluripotent profile (3.1) to the stem-cell data

| Gene names | $\hat{\gamma}_1$ | $\hat{\gamma}_2$ | $\hat{\gamma}_3$ | $U$   |
|------------|------------------|------------------|------------------|-------|
| Oct4       | 0.48             | 2.77             | 0.20             | 2.53  |
| Utf1       | 0.54             | 1.82             | -0.42            | 2.46  |
| Tdgfl      | 1.04             | 1.88             | 0.22             | 1.80  |
| S1c35f2    | 0.32             | 1.69             | -0.17            | 1.53  |
| Trh        | 0.44             | 1.71             | -0.69            | 1.50  |
| Foxd3      | 0.14             | 1.79             | -0.17            | 1.33  |
| Musd1      | 0.15             | 2.00             | -0.62            | 1.17  |
| Skil       | 0.15             | 1.66             | -0.83            | 1.16  |
| Pou6f1     | 0.54             | 1.66             | 0.24             | 1.13  |
| Par2       | 0.33             | 1.58             | 0.60             | 0.75  |
| Nanog      | 0.31             | 1.99             | 0.88             | 0.69  |
| S1c7a3     | 0.09             | 2.45             | -0.58            | 0.67  |
| Gng3       | 0.15             | 1.55             | -0.42            | 0.33  |
| Skil       | 0.23             | 1.54             | -0.74            | 0.28  |
| Rae-28     | 0.14             | 1.51             | -0.29            | 0.08  |

Fig. 2. Fitted log ratios with respect to day 0 for the ranked genes for the pluripotency profile (3.1).
Figure 2 shows the fitted log ratios with respect to day 0 for the 4 time points: day 0, day 3, day 6, and day 9. Therefore, all the profiles will pass through zero on day 0. The profiles demonstrate the required trajectory: equal expression for day 0 and day 3, higher gene expression levels for days 0 and 3 compared to day 9, and the gene expression level for day 6 lying between the gene expression levels for days 0 and 3 and that for day 9.

The top ranked gene, Oct4, is well known to be associated with pluripotency (Rodda and others, 2005; Loh and others, 2006) and would therefore be expected to appear among the top ranked genes for pluripotency in this experiment. Other genes of note in Table 1 are Utf1 (rank 2), which is associated with undifferentiated embryonic cell transcription (Nishimoto and others, 2005), and Nanog (rank 11), which is central to embryonic stem-cell pluripotency (Wang and others, 2006).

The recent article by Wang and others (2006) isolated proteins associated with the protein Nanog and thus with pluripotency. Of the 38 proteins discussed in Wang and others (2006), Oct4 and Nanog appeared in our list of ranked genes using model (3.1) with ranks 1 and 11, respectively. The remaining proteins were not in the ranked genes as the profiles of the associated mRNAs are not consistent with profile (3.1).

As discussed previously, the choice of the neighborhood around zero assumed for equivalence (i.e. $\epsilon$) should be decided upon in consultation with biologists. However, this is problematic since biologists still have relatively little explicit knowledge of genewise expression variability.

To investigate the potential effects of altering the neighborhood defined by $\epsilon$, the primary analysis was repeated assuming values of $\epsilon = 0.5$, 1, 1.5, and 2. In Figure 3, the profiles for the genes which have observed profiles that lie within the rejection region are plotted for each choice of equivalence neighborhood. As the equivalence neighborhood width ($\epsilon$) increases, more genes have profiles that lie within the rejection region, but there is greater variation between the gene expression levels for day 0 and day 3. Nevertheless, gene profiling in this application has been demonstrated to be reasonably robust. For $\epsilon = 0.5$, 1, and 1.5, Oct4 was ranked as the top gene, while for $\epsilon = 2$, it had only dropped to rank 2.

It is well known that the gene Sox2 is commonly associated with pluripotency (Rodda and others, 2005), but it was not in the ranked genes using the gene expression profile (3.1) because the fitted gene expression profile for Sox2 is very different from the pluripotent profile used in the analysis. The criteria for the profile of Sox2 are higher gene expression level on day 0 compared to the gene expression levels for days 6 and 9; equivalent gene expression levels on days 6 and 9; and the gene expression level for day 3 to lie between the gene expression level for day 0 and the levels for days 6 and 9. Gene profiling can be used to rank the genes according to these alternative criteria. An appropriate model for Sox2 is

$$
\mu = \begin{pmatrix}
1 & 1 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1/2 \\
1 & 0 & 0 & -1/2
\end{pmatrix} \gamma,
$$

in which $\gamma_0$ is unrestrained, $\gamma_1 > 0$, $\gamma_2 > 0$, and $\gamma_3$ is equivalent to zero. This model was fitted to the data and the ranked genes are shown in Figure 4(a). The ranked genes were Cpt1a, 1200014E20Rik, 2210409E12Rik, Sox-2, Np-1, Birc5, 5730419I09Rik, MGI:1922156, retSDR3, and clone RP21-505L19 on chromosome 5. Sox2 was ranked at position 4. Although retSDR3 has the same desired form with the largest apparent magnitude, as shown in Figure 4(a), it is only ranked ninth with respect to genes matching the Sox2 profile. The low ranking results from the large gene expression variance (0.181) of retSDR3 relative to the other ranked genes (average gene expression variance of 0.054). This illustrates that if 2 genes have the same coefficient values, gene profiling will rank lower the gene which has the larger variance and thus more uncertainty about its true profile.

As mentioned in Section 1, we also considered both Pareto optimization and the methodology discussed in Lönntstedt and others (2001) to identify genes associated with pluripotency. Pareto optimization
Fig. 3. Fitted log ratios with respect to day 0 for the ranked genes with (a) $\epsilon = 0.5$, (b) $\epsilon = 1$, (c) $\epsilon = 1.5$, and (d) $\epsilon = 2$.

Fig. 4. (a) Fitted log ratios with respect to day 0 for the top 10 ranked genes for the Sox2 profile. (b) Fitted log ratios with respect to day 0 for the Pareto-optimal genes for the Sox2 profile.
Gene profiling for determining pluripotent genes

considers genes which are Pareto optimal with respect to a suitable set of criteria (Hero and Fleury, 2004; Fleury and others, 2002) and requires the profiles simultaneously to obey several constraints as closely as possible. However, although Pareto optimality is formulated in terms of multiple constraints, a gene that is optimal for one constraint will be Pareto optimal irrespective of how poorly it performs on all other constraints. For comparison with gene profiling, Pareto optimization was applied to the pluripotent data with the Sox2 profile. The criteria chosen to find genes consistent with this profile were

1. minimize \(|\mu_9 - \mu_6|\),
2. minimize \((\mu_6 + \mu_9)/2 - \mu_3\), and
3. minimize \(\mu_3 - \mu_0\),

where \(\mu_i\) is the estimated mean log ratio with respect to day 0 for day \(i\), \(i = 1, 3, 6, \) and 9. There were 25 Pareto-optimal genes, whose profiles are shown in Figure 4(b). These genes do not match the desired profile as closely and as consistently as those found by gene profiling (Figure 2). Of the 10 genes identified by gene profiling, only 5 appear in the Pareto-optimal gene list. Moreover, the Sox2 gene itself is not included in the Pareto-optimal set.

Lönnstedt and others (2001) used the inner product between a gene’s expression levels and a prespecified profile to rank the genes as described in Section 1. The difficulty with this method is that, without suitable standardization, a large inner product is not a necessary or sufficient condition for close concordance between \(y\) and \(u\). We applied this approach to the Sox2 profile, for which the observed mean log ratios calculated from the pluripotent data are \((0, -0.5033, -0.7927, -0.7979)\). Therefore, we set \(u = (0, -0.5033, -0.7927, -0.7979)\) and ranked the genes by the inner product of their observed mean log ratios and \(u\). In Lönnstedt and others (2001), a Q-Q plot of the inner product was used to choose a cutoff point for genes of potential interest. The inner products are expected to lie roughly on a straight line, and any genes deviating from the line are of interest. For the Sox2 profile, we chose a cutoff point of 4 which resulted in 37 genes of interest, whose profiles are shown in Figure 5. Again, the genes do not match the required profile as closely or consistently as those identified by gene profiling. The target gene, Sox2, is ranked 454 by the Lönnstedt method. Moreover, 9 of the 10 genes found by gene profiling were
not ranked within the top 200 using the Lönstedt approach. Although the Lönstedt method did rank the gene \textit{retSDR3} first, we contend that this high rank is primarily due to the gene’s large relative variance, as noted in Section 5.

## 5. Discussion of Further Work

In general, gene profiles of interest to molecular biologists often consist of 2 types of criteria: equal and different gene expression at different time points. Gene profiling provides a straightforward methodology to filter genes which satisfy these 2 types of criteria simultaneously. We believe that this has not been accomplished using previously available techniques. By simultaneously testing for all criteria, gene profiling effectively filters out and excludes genes that are only partially consistent with the required profile. We now touch on some areas requiring further work.

As noted in Section 3.3, to test for a parameter being equal to zero, a neighborhood of width \( \epsilon \) is defined. This neighborhood is the amount that the parameter could vary and still be considered equivalent to zero. In this paper, the choice of \( \epsilon \) was based on plotting profiles for the various choices of \( \epsilon \) and choosing the best \( \epsilon \) to give the required prespecified profiles. Ideally, the choice of \( \epsilon \) should be based on consultation with biologists, to the extent that such knowledge is available. We anticipate that such knowledge will gradually accrue over time, as microarray and other new genomics technologies are more widely applied in molecular biology and genetics.

Another area requiring further research is the invariance (or otherwise) of reparameterization. Wellek (2002) notes:

“... in contrast to the corresponding conventional testing problems with the common boundary of null and alternative hypothesis [sic] being given by zero, equivalence problems remain generally not invariant under redefinitions of the main parameter.”

To illustrate this point, consider the problem of finding marker genes for day 3 in the stem-cell experiment. The criteria for such genes are high gene expression level on day 3, as well as equal and low gene expression levels on day 0, day 6, and day 9. Examination of the profile reveals 3 possible models:

\[
\begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 1 & -\frac{1}{3} & \frac{1}{3} \\
1 & 0 & 0 & -1 \\
1 & 0 & -1 & 0
\end{pmatrix} \gamma, \quad \begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 1 & -\frac{2}{3} & \frac{1}{3} \\
1 & 0 & -1 & 1 \\
1 & 0 & -1 & 0
\end{pmatrix} \gamma, \quad \begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 1 & -\frac{2}{3} & -\frac{1}{3} \\
1 & 0 & -1 & 0 \\
1 & 0 & -1 & -1
\end{pmatrix} \gamma,
\]

where \( \gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)' \) with \( \gamma_0 \) unrestrained, \( \gamma_1 \) significantly positive, \( \gamma_2 \) equivalent to zero, and \( \gamma_3 \) equivalent to zero.

The 3 models may not necessarily give the same results. This is because equivalence is not transitive, that is, if \( \mu_0 \) is equivalent to \( \mu_6 \) and \( \mu_6 \) is equivalent to \( \mu_9 \), it is not necessarily true that \( \mu_0 \) is equivalent to \( \mu_9 \). This is because equivalence is defined in a neighborhood and so a “drift” resulting in \( \mu_0 \) and \( \mu_9 \) being too far apart to be considered equivalent can occur. Methods to impose invariance are currently under investigation by the authors. Although this is an interesting area of research, invariance of reparameterization is not unique to gene profiling, and in many cases the research hypotheses will dictate the optimal model to use.

Gene profiling can also be used to select genes that have a similar profile to an existing profile. There are 2 possible ways to achieve this. First, we can obtain the vector, \( u \), of the mean gene expression levels for the existing profile. Then, we can use the following linear model for the true mean gene expression
levels, $\mu$:

$$\mu = \gamma_0 1 + \gamma_1 u + \sum_{i=2}^{p} \gamma_i v_i,$$

where $v_i, i = 2, \ldots, p$, are orthogonal to $1$ and $u$, and $p$ is the number of time points. Similar genes are identified by utilizing gene profiling with the constraints: $\gamma_0$ unconstrained, $\gamma_1 > 0$, and $\gamma_i = 0$, for $i = 2, \ldots, p$.

The second method to find similar genes is as follows. The existing profile is plotted and the salient relationships identified. The relationships to consider are the inequalities and equalities of gene expression between the time points. Once the predominant relationships are identified, then the basis vectors can be chosen in a similar fashion to the methods used for the pluripotent example in Section 3.1 and the Sox2 profile in Section 4. An algorithm to calculate the appropriate basis vectors is contained in Section C of the supplementary material available at Biostatistics online (http://www.biostatistics.oxfordjournals.org).

Gene profiling is easily implemented by fitting the model to the data using limma and then calculating the $U$ statistics. The calculation of the $U$ statistics was written in C to decrease the run time but is easy to implement in R.

To conclude, gene profiling introduces a flexible method to select genes for a prespecified time course profile. Gene profiling is straightforward to implement in practice, requiring only small modifications to the R package limma, and can be used to select for most profiles of interest to biologists. Gene profiles can be used for any number of time points. Beyond say 10 or 12 points, however, one might start to consider methods appropriate for time series data. The application of gene profiling in this article has been to 2-color microarrays, but it could readily be modified for use for other microarrays platforms, such as Affymetrix GeneChip (Lockhart and others, 1996), and for other technologies where it is required to rank observations by correspondence with a prespecified profile.

ACKNOWLEDGMENTS

We thank the Rathjen group for the use of the stem-cell data and Dr Chris Wilkinson and Professor Terry Speed for useful information about stem-cell experiments. Conflict of Interest: None declared.

FUNDING

The Australian Research Council (Discovery Project Grant [0450632] to G.F.V.G and P.J.S); University of Adelaide (George Fraser PhD scholarship to J.T.).

REFERENCES

Adams, R. and Bischof, L. (1994). Seeded region growing. IEEE Transactions on Pattern Analysis and Machine Intelligence 16, 641–647.

Ahnert, S. E., Willbrand, K., Brown, F. C. S. and Fink, T. M. A. (2006). Unbiased pattern detection in microarray data series. Bioinformatics 22, 1471–1476.

Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T. S. and Simon, I. (2003). Continuous representations of time-series gene expression data. Journal of Computational Biology 10, 341–356.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. Technometrics 24, 295–300.

Brown, A. L., Wilkinson, C. R., Waterman, S. R., Kok, C. H., Salerno, D. G., Diakiw, S. M., Reynolds, B., Scott, H. S., Tsykin, A., Glonek, G. F. and others (2006). Genetic regulators of
myelopoiesis and leukemic signaling identified by gene profiling and linear modeling. *Journal of Leukocyte Biology* **80**, 1–15.

CELEUX, G., MARTIN, O. AND LAVERGNE, C. (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling* **5**, 243–267.

D’AMOUR, K. A. AND GAGE, F. H. (2003). Genetic and functional differences between multipotent neural and pluripotent embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 11866–11872.

DUDOIT, S., YANG, Y. H., CALLOW, M. J. AND SPEED, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. AND BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863–14868.

ERNST, J., NAU, G. J. AND BAR-JOSEPH, Z. (2005). Clustering short time series gene expression data. *Bioinformatics* **21**, 159–168.

FLEURY, G., HERO, A., YOSHIDA, S., CARTER, T., BARLOW, C. AND SWAROOP, A. (2002). Pareto analysis for gene filtering in microarray experiments. *Proceedings of the 11th European Signal Processing Conference*. 3, 165–168.

GHOSH, D. AND CHINNAIYAN, A. M. (2002). Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* **18**, 275–286.

GLONEK, G. F. AND SOLOMON, P. J. (2004). Factorial and time course designs for cDNA microarray experiments. *Biostatistics* **5**, 89–111.

HERO, A. O. AND FLEURY, G. (2004). Pareto-optimal methods for gene ranking. *The Journal of VLSI Signal Processing* **38**, 259–275.

LOCKHART, D. J., DONG, H., BYRNE, M. C., FOLLETTE, M. T., GALLO, M. V., CHEE, M. S., MITTMANN, M., WANG, C., KOBAYASHI, M.,HORTON, H. and others (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**, 1675–1680.

LOH, Y.-H., WU, Q., CHEW, J.-L., VEGA, V. B., ZHANG, W., CHEN, X., BOURQUE, G., GEORGE, J., LEONG, B., LIU, J. and others (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics* **38**, 431–440.

LÖNNSTEDT, I., GRANT, S., BEGLEY, G. AND SPEED, T. P. (2001). Microarray analysis of two interacting treatments: a linear model and trends in expression over time. *Technical Report 7/2001*. Uppsala, Sweden: Department of Mathematics - Uppsala University.

MCLACHLAN, G. J., BEAN, R. W. AND JONES, L. B. T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* **22**, 1608–1615.

NGUYEN, D. V., ARPAT, A. B., WANG, N. Y. AND CARROLL, R. J. (2002). DNA microarray experiments: biological and technological aspects. *Biometrics* **58**, 701–717.

NISHIMOTO, M., MIYAGI, S., YAMAGISHI, T., SAKAGUCHI, T., NIWA, H., MURAMATSU, M. AND OKUDA, A. (2005). Oct-3/4 maintains the proliferative embryonic stem cell state via specific binding to a variant octamer sequence in the regulatory region of the UTF1 locus. *Molecular and Cellular Biology* **25**, 5084–5094.

RAMALHO-SANTOS, M., YOON, S., MATSUZAKI, Y., MULLIGAN, R. C. AND MELTON, D. A. (2002). “Stemness”: transcriptional profiling of embryonic and adult stem cells. *Science* **298**, 597–600.

R DEVELOPMENT CORE TEAM (2006). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

RODDA, D. J., CHEW, J.-L., LIM, L.-H., LOH, Y.-H., WANG, B., NG, H.-H. AND ROBSON, P. (2005). Transcriptional regulation of Nanog by Oct4 and Sox2. *The Journal of Biological Chemistry* **280**, 24731–24737.

SMYTH, G., YANG, Y. AND SPEED, T. (2003). Statistical issues in cDNA microarray data analysis. *Methods in Molecular Biology* **224**, 111–136.
Gene profiling for determining pluripotent genes

SMYTH, G. K. (2005). *Limma: Linear Models for Microarray Data*. New York: Springer, pp. 397–420.

TAI, C. Y. AND SPEED, T. P. (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Annals of Statistics* **34**(5), 2387–2412.

TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S., DIMITROVSKY, E., LANDER, E. S. AND GOLUB, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2907–2912.

TAVAZOIE, S., HUGHES, J. D., CAMPBELL, M. J., CHO, R. J. AND CHURCH, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**, 281–285.

WANG, J., RAO, S., CHU, J., SHEN, X., LEVASSEUR, D. N., THEUNISSEN, T. W. AND ORKIN, S. H. (2006). A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444**, 364–368.

WELLEK, S. (2002). *Testing Statistical Hypotheses of Equivalence*. Boca Raton, FL: CRC Press.

YANG, Y., BUCKLEY, M. AND SPEED, T. (2001). Analysis of cDNA microarray images. *Briefings in Bioinformatics* **2**, 341–349.

YANG, Y. H. AND SPEED, T. (2003). Design and analysis of comparative microarray experiments. In: Speed, T. (editor), *Statistical Analysis of Gene Expression Microarray Data*. Chap. 2. Boca Raton, FL: Chapman & Hall, pp. 35–91.

YEUNG, K. Y., FRALEY, C., MURUA, A., RAFTERY, A. E. AND RUZZO, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987.

[Received July 31, 2007; first revision January 23, 2008; second revision February 25, 2008; third revision March 26, 2008; accepted for publication May 12, 2008]