Sample heterogeneity severely limits the conclusions that can be made about specificity of gene expression and may explain in part why the results of numerous gene expression experiments have failed rigorous validation (Michiels et al., 2005). Given a heterogeneous sample there exists the potential for identifying disease biomarkers and improving our understanding of cellular processes (Pittman et al., 2006). In breast cancer research and publicly available experimental datasets found in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) repository.

We present a method for deconvoluting expression from a heterogeneous sample into components that reflect the contributions to the observed expression attributable to each component or tissue type. The key component of this method is the estimation, from a mixed tissue sample, of the proportion of mRNA from a tissue type. The proportion of each component in a mixed tissue sample; this estimate can be used to provide estimates of gene expression from each component. We demonstrate our technique on xenograft samples from breast cancer research and publicly available experimental datasets found in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo; Barrett et al., 2006).

2 APPROACH

Several approaches have been taken to the problem of expression deconvolution and each approach depends on access to different types of information, different statistical assumptions and different objectives.

If there are genes known to be expressed exclusively in one tissue type, then these genes can be used to estimate the proportion of expression coming from that tissue. For example, the program DECONVOLUTE (Lu et al., 2003) uses simulated annealing and genes expressed only during specific cell cycles to identify the proportion of cells in each cycle from an asynchronous cellular population (Causton et al., 2003). In breast cancer Cleator et al. (2006) noticed that the proportion of benign tissue of biopsy samples can significantly affect expression profiles, and taking into consideration this proportion can improve response prediction.

© The Author(s) 2010. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
of tissue-specific genes. Note that low specificity of microarray hybridizations has been suggested to be one of the prime measures affecting discrepancies in gene-expression profiles between different probes targeting the same region of a given transcript or between different microarray platforms (Kohlai and Weingarten-Baror, 2008). We do not assume knowledge of cell- or tissue-specific genes in our method, although such knowledge may be available, particularly for samples from xenograft studies (where the tissues of interest are from different species).

Similarly, several researchers have used expression data from purified reference tissue types to determine the expression of each tissue type in heterogeneous samples (Lahdesmaki et al., 2005; Venet et al., 2001). For example, Wang et al. (2006) use a method similar to that of Lu et al. (2003), mentioned above, to determine the proportions of each cell type in a mixed sample. This method generates estimates by obtaining solutions to linear equations via simulated annealing. These approaches depend on having expression data from a purified reference sample for each cell or tissue type, which may not be available.

Another approach uses proportions of each sample or cell type, assessed by pathologists, to establish either tissue-specific expression or differential expression between mixed and control samples. In Stuart et al. (2004) linear regression models, regressing expression on fractional content of tumor (or stroma), were used to estimate the expected cell-type expression as the regression coefficient. A more sophisticated statistical approach was used by Ghosh (2004) to determine differential expression in the presence of mixed cell populations. In his approach, a pathologist’s assessments of the proportions of each cell type were used in a hierarchical mixture model to model the data. A combination of methods of moments procedures and the expectation-minimization (EM) algorithm provided estimates of the model parameters. Although not shown in the publication, this method could be adapted to provide expression estimates specific to each cell type, as opposed to estimates of differential expression. Unfortunately, the assessment of a pathologist only provides the proportion of each cell or tissue type in the sample, and not an assessment of the amount of mRNA or protein attributable to each. It is well known that the total amount of mRNA generated by tumor cells, for example, is much higher than the amount generated by normal cells. As a result estimates of expression based on pathological assessments of tissue proportions may not be accurate.

Finally, an approach exists to use expression data from a single cell type to determine the proportion of each cell type in a heterogeneous sample (Gosink et al., 2007). This method depends on the estimation of the minimum of a proportion, a minimum that provides a good estimate in noiseless or simulated data. However, this minimum is much more difficult to estimate in noisy data, and microarray data is inherently noisy. Our research builds upon this work by providing a method for estimating this minimum that has reasonable accuracy and can be applied in situations where one or multiple heterologous samples are available.

3 METHODS

First, we will discuss the idea of estimating the proportion of a single cell or tissue type in a two-type mixed sample. We will then describe the role of data transformation in this estimation and the interest in finding the point of minimum radius of curvature. Finally, we will describe the use of the bootstrap (Efron, 1979) for obtaining a standard error for our estimate.

3.1 Proportion of tumor as a minimum ratio

The idea of estimating the proportion of one type in a two-type mixed sample comes from Gosink et al. (2007). As they describe, let a be a purified sample of one type and AB a mixed sample, composed of tissue or cell types A and B. Let \( E_i(A|B) \) be expression of gene i in Sample AB(A) for \( i = 1, \ldots, m \). Let \( E_i(AB) = \{E_i(AB_1), \ldots, E_i(AB_m)\} \). We want to estimate \( p_A \), the proportion of expression in the mixed sample (Sample AB) due to tissue type A. For a given gene \( i \) we can express \( E_i(AB) \) as

\[
E_i(AB) = p_A E_i(A) + (1 - p_A) E_i(B).
\]

Let \( R_i = R_{mix,type} = E_i(AB)/E_i(A) \). In the noiseless case,

\[
R_i = p_A E_i(A)/E_i(A) + (1 - p_A) E_i(B)/E_i(A).
\]

Note that for a fixed \( p_A \), this ratio is at its minimum when \( E_i(B) = 0 \), since expression is assumed to be non-negative. Hence, if \( E_i(0) > 0 \),

\[
\lim_{E_i(B) \rightarrow 0} R_i = p_A + (1 - p_A) E_i(0)/E_i(A).
\]

Thus, under the assumption that \( E_i(B) \rightarrow 0 \) for some sequence of \( i \),

\[
\min_i R_i = p_A + (1 - p_A) E_i(0)/E_i(A).
\]

This can be seen in Figure 1 where rank-sorted ratios \( R_i \) have been plotted from 'electronic' simulated data at a range of proportion values \( p_A \). The 'electronic data' was generated by computationally combining expression values from purified samples of each composite type in these specific proportions, e.g. for \( p_A = 0.25 \) the electronic data is \( 0.25 E_i(A) \rightarrow 0.75 E_i(B) \) where \( E_i(A) \) is expression values from a purified sample of breast cancer cell mRNA and \( E_i(B) \) the expression values from a purified sample of normal mouse lung mRNA. Note that the values of \( R_i \) are sorted from lowest to highest.

Unfortunately, the minimum ratio is an underestimate of the true proportion value \( p_A \) for simulated noisy data and for observed data (Gosink et al. 2007). For example, Figure 2 shows observed data from a titration series \( p_A = 0.25, 0.5 \) and 0.75 of breast cancer cell mRNA (MDA231) and normal mouse lung mRNA. By a titration series, we mean a set of mixed samples (breast cancer and normal lung) in which each sample...
A transformation that increases small ratio values while shrinking larger ratio values may improve the accuracy of this estimate. To explore this proposition, we want to find the value of \( \alpha \) as a function of arc length \( \alpha \), i.e. within each dataset and across datasets the value of \( \alpha \) is dependent on \( \alpha \). However, this function depends on \( \alpha \), the value we are trying to estimate.

We acknowledge that the minimum value of \( \alpha \) is sensitive to the noise in the data, particularly in relation to the mean or the median. Hence we decided to explore the possibility of using information from a summary statistic of \( \alpha \), e.g. mean or median) as a function of \( \alpha \) to determine the correct value of \( \alpha \) and hence the value of our estimate \( \alpha \). The mean of \( \alpha \) is defined as a function of \( \alpha \) as follows:

\[
\alpha = \frac{1}{m} \sum_{i=1}^{m} \log \left( 1 + a \log \left( A_i \right) \right)
\]

where \( m \) is the number of expression values (i.e., number of genes). The median is analogously defined. We decided to plot the mean and median of \( \alpha \) for a fixed \( a \) across a range of values of \( a \). Example plots for two different titration series (MDA231/mouse lung at \( pA = 0.5 \) and data from MicroArray Quality Control (MAQC) project 2006 (GEO accession 5350)) at \( pA = 0.75 \) are shown in Figure 3. The value of \( a \) that provides the most accurate estimate of \( pA \) is marked by a vertical black line.

The value of \( a \) that provides the most accurate estimate of \( pA \) in these plots and many others, is located at what one may refer to as the ‘knee’ or ‘elbow’ of the curve. This point may be familiar from principal components analysis as the point on a scree plot that indicates the number of significant principal components (Jolliffe, 2002). To calculate this point, we need a mathematical definition for the ‘elbow’ of a curve.

### 3.3 Minimum radius of curvature

We want to find the value of \( \alpha \) at the ‘elbow’ of the curve defined by \( \alpha \) as a function of \( \alpha \). The ‘elbow’ of a curve is the point at which the tradeoff between pulling low values up and pulling high values down (values of \( \alpha \)) is optimal. Here, we formalize this by choosing that point at which the radius of curvature is at its minimum. The radius of curvature \( \rho(x) \) is defined as the inverse of the vector norm of the second derivative of the curve, expressed as a function of arc length \( \alpha \), i.e.

\[
\rho(x) = \frac{1}{\left\| C'(x) \right\|}
\]

where \( C \) is the curve of interest originally parameterized in terms of \( x \) (Lipschutz, 1969). Thus, to find the value of \( \alpha \) of interest several steps are required. First, we need to represent the function \( \rho(x) \) as a curve in the plane. Second, we must reparameterize this curve in terms of arc length \( s \). Third, we use the reparameterized curve to determine the value of arc length \( \alpha \) that minimizes the radius of curvature \( \rho(x) \). Finally, we determine the value of \( \alpha \) that corresponds to \( s \).

#### 3.3.1 Radius of curvature in terms of arc length

Recall that a parameterized curve in the plane is of the form

\[
s(x) = x_1(x) e_1 + x_2(x) e_2, \quad x \in [\alpha, \alpha']
\]

where \( x_1, x_2 \) are the coordinate functions, \( e_1 \) and \( e_2 \) are the natural basis and \( \alpha \) is the parameter of the curve. To define the radius of curvature of \( s(x) \) at a point \( x \), we first reparameterize in terms of arc length \( s \). The arc
length parameterization is defined to be the parameterization with unit speed along the curve. This eliminates the possibility of an unnaturally high or low radius of curvature simply due to the local speed of traversal of the curve.

The arc length $s$ of a curve is defined as

$$s(a) = \int_a^b |ds/du| du$$

(3)

where $| |$ is the Euclidean norm. Now consider a function $f(u)$ and observe that its graph $(u, f(u))$ is a geometric curve in the plane. Thus, as in Equation (2), we can write

$$s(a) = (a, u) + f(u) \overline{a}, \quad u \in \left[a', a''\right].$$

Hence

$$ds/du = 1 + f'(u)^2 \overline{a}, \quad |ds/du| = 1 + f'(u)^2.$$

So the arc length parameter (Equation [3]) is given in terms of $u$ by

$$s(u) = \int_a^u \sqrt{1 + f'(u)^2} \overline{a}, \quad a \in \left[a', a''\right].$$

Since the parameterization is in terms of unit speed, it is invertible, so we can write $u = u(a)$ as well. Thus, $s' = s(x'$) and $s'' = s(x''$). The radius of curvature of a geometric curve $C$ as stated in Equation (1) can now be defined as

$$\rho(u) = 1/\left|f''(u) \overline{a}\right|$$

for the curve $C = (u, \overline{a}(u))$. We will argue that choosing $a$ to minimize $\rho(u)$ leads to a good estimate of $p_a$ over $[a', a'']$.

3.3.2 Implementation Given the definitions in the last subsection, it remains to obtain the arc length parameterization for the curve $(u, \overline{a}(u))$ and find the value of $a$ that corresponds to $s'$. Replacing $f(u)$ with $\overline{a}(u)$ we have the following:

Partition the interval $[a', a'']$ uniformly by setting

$$a' = u_1 < u_2 < \cdots < u_k = a''$$

and $u_i - u_{i-1} = a'' - a'$. Then,

$$s(u) = \int_a^{u_i} \sqrt{1 + f'(u)^2} \overline{a} du = \frac{1}{k} \left(1 + |\overline{a}||\overline{a}_{i-1}|\right)$$

where we approximate $\overline{a}(u)$ as

$$\overline{a}(u) = \frac{\overline{a}_i + \overline{a}_{i-1}}{2}.$$

This will yield a one-to-one relationship between $u$ and $s$, hence a one-to-one relationship between $s$ and $\overline{a}(u)$. Once we have this we can find the value of $a$ that minimizes the radius of curvature $\rho(u)$, i.e. maximizes $\left|f''(u) \overline{a}(u)\right|$ over $x'$. $x''$.

3.3.3 Determining $s'$ and $a$ To find the maximum of

$$\left|f''(u) \overline{a}(u)\right| = \sqrt{\left(\overline{a}(u)\right)^2 - \left(\overline{a}(u)\right)^2},$$

we find the value of $a$, $s'$, that maximizes the absolute value of the second derivative with respect to $x$ using centered difference approximations (Ames, 1977). Approximate $\overline{a}(u)$ by

$$\overline{a}(u_{i+1}) - 2\overline{a}(u_i) + \overline{a}(u_{i-1})$$

and $\left|f''(u) \overline{a}(u)\right|$ over a range of values $[x', x'']$ and determine the value $s'$ that minimizes $\overline{a}(u)$. Note that this method for finding $s'$ (and subsequently $a$) only works if the two axes of the plot for $\overline{a}(u)$ are similarly scaled. If the two scales are not equal, they must be equalized prior to calculating $s'$ by rescaling one axis to be the same length as the other. For example, to rescale the axis for $\overline{a}(u)$ we would use values of the following in place of $\overline{a}(u)$

$$(\max(a) - \min(a)))/(\max(\overline{a}(u)) - \min(\overline{a}(u)))$$

That is, the range of the function $\overline{a}(u)$ is the same as the range of the parameter $u$. This ‘scaling’, like the arc length parameterization, seems necessary to prevent arbitrary choices from dominating the solution.

One key task is choosing $k$ large enough so that the approximation of the second derivative with respect to second differences is accurate over the range $[a_0, a_T]$. We found that $k$ of several thousand worked well in the examples in Section 4.

3.4 Bootstrap estimates of standard error

We used a simple bootstrap resampling procedure (Efron, 1979) to generate standard errors for our estimate of $p_a$. For a given dataset of $n$ observations and $m$ genes, we draw $T$ bootstrap samples, each sample contains expression values of $m'$ genes drawn at random with replacement where $m' = 0.6 + m$ (so a total of $m'$ values). From each sample $j = 1, \ldots, T$, we calculate the mean of $\overline{a}(u)$ across values of $u$ in a given range. We then determine the value of $u$ that corresponds to the minimum radius of curvature $(s')$ of $\overline{a}(u)$, plotted as a function of $u$ (as described in Section 3.3). This value of $u$ is used to generate $\overline{a}(u)$ for the genes in sample $j$ and determine its minimum, i.e. our estimate of $p_a$. The result of our bootstrap procedure is $T$ estimates of $p_a$, $(\hat{p}_a_1, \ldots, \hat{p}_a_T)$, one for each sample. The SD of these estimates is taken as the standard error of our estimate of $p_a$, and a $(1 - \alpha)$ confidence interval for our estimate is $(\hat{p}_a^-_T, \hat{p}_a^+_T)$ where $\hat{p}_a^-_T$ and $\hat{p}_a^+_T$ are the $(\alpha/2)$th and $(1 - (\alpha/2))$th percentiles of our 100 estimates of $p_a$.

Stated as pseudo-code for clarity, our procedure is as follows:

(1) Generate $T$ bootstrap samples where each sample contains $m'$ expression values, i.e. expression values for $m'$ genes from each sample. The $m'$ genes, $m' = 0.6 + m$, are selected at random and with replacement.

(2) For each sample, calculate the values of $\overline{a}$, $i = 1, \ldots, m'$, for a range of values of $u$.

(3) For each sample, calculate the values on the curve $(a, \overline{a}(u))$ for a range of values of $u$, using the result of step 2.

(4) For each sample, use the curve calculated in Step 3 to determine the value of $u$ that corresponds to the minimum radius of curvature $(s')$ (as described in Section 3.3); label this value as $\alpha_j$ for each sample $j = 1, \ldots, T$.

(5) For each sample $j$, use the values of $\overline{a}_j$ that correspond to $\alpha_j$ (as calculated in Step 2) and determine its minimum, i.e. our estimate of $p_a$. This yields $(\hat{p}_a_1, \ldots, \hat{p}_a_T)$.

(6) Calculate the standard error of our estimate (as the SD of $(\hat{p}_a_1, \ldots, \hat{p}_a_T)$) and a $(1 - \alpha)$ confidence interval as $(\hat{p}_a^-_T, \hat{p}_a^+_T)$ where $\hat{p}_a^-_T$ and $\hat{p}_a^+_T$ are the $(\alpha/2)$th and $(1 - (\alpha/2))$th percentiles of our 100 estimates of $p_a$.

As a sort of stability analysis, we choose a range of values for $T$ in our computations below to see whether there was any obvious relationship between the size of $T$ and the likelihood that a bootstrapped confidence interval contained the true value. The results in Table 2 suggest that the size of $T$ and the accuracy of the bootstrap intervals is slight at most.

4 RESULTS

We implemented our procedure for estimating $p_a$ in several gene expression datasets, both proprietary and public, in which expression data was generated from samples composed of two tissue/cell types. Some of the samples consist of different cell types from the same
organism, while other samples are a mix of cell types from different organisms. The proportion of each component type is known, as the data come from titration series; we use these values to assess the accuracy of our estimates.

### 4.1 Data

Our data consists of six datasets obtained either from the University of Miami School of Medicine (UMiami) or the NCBI GEO (Barrett et al., 2006). The UMiami datasets were created as a titration series of RNA from breast cancer cells (either MDA231 or MCF7) and normal mouse lung cells. The expression platform is Illumina Human WG-6 version 2 (MCF7) or version 3 (MDA231) (Illumina Inc., 2009); chips were processed at two different laboratories. The data from GEO includes titration series of Universal Human Reference RNA and Human Brain RNA from the MAQC study (MAQC Consortium, 2006). We selected data processed at two different laboratories and on two different platforms, either Human-6 BeadChip 48K version 1 (Illumina Inc., 2009) or HG-U133 Plus 2.0 GeneChip (Affymetrix Inc., 2009). Two other datasets from GEO were also included in our studies; these data include two titration series of mouse T and B cells (Shearstone et al., 2006). These sets were processed on the Mouse 430A version 2 GeneChip platform (Affymetrix Inc., 2009). The details of each dataset are presented in Table 1.

The method of normalization of gene expression data can impact substantially which probes are identified as detected and which probes are differentially expressed between conditions (Dunning et al., 2008b; Johnstone et al., 2008). For this reason, we implemented several normalization methods on our proprietary datasets, while using the available normalized data for the publicly available datasets. The normalization methods for the Illumina data include quantile normalization and spline normalization as implemented in the R package beadarray (Dunning et al., 2008a; R Development Core Team, 2009) and cubic normalization as implemented in the Illumina BeadStudio software (Illumina Inc., 2009). After normalization, only those genes with a detection P-value <0.01 in all samples (Illumina Inc., 2009) or considered present in all samples according to the Affymetrix MAS5 algorithm (Affymetrix Inc., 2009) were included in further analyses (i.e. bootstrap estimation of $p_A$ by the procedure described in Section 3.4).

### 4.2 Accuracy of estimation

Select results of our bootstrap estimation procedure for each dataset are shown in Figure 4. For the UMiami datasets, we chose to display results for only one normalization method for brevity.

In ~90% of cases, our point estimate is within 5% of the true proportion; in ~80% of cases, the 90% bootstrap confidence interval for our estimate contains the true value of $p_A$. We note that our method found the BIIB 100 dataset to be the most challenging. This is no surprise as this titration series was designed with very low levels of mRNA, as a challenge to the procedure used for RNA amplification prior to running the expression assay (Shearstone et al., 2006). In other words, this data was generated from a very small amount of biological material so the estimation of the proportion of the biological components is very challenging. There is evidence in Figure 4 of an interaction between the normalization procedure and the accuracy of our estimation method. For example, we tend to overestimate $p_A$ when the data is quantile normalized. We note that our underlying model has some level of uncertainty, the stated level of confidence is an overestimate of the actual level of confidence. In other words, model uncertainty...

### Table 1. Available datasets

| Source      | Type       | Platform | Proportion | n | Norm | GEO       |
|-------------|------------|----------|------------|---|------|-----------|
| UMiami      | MDA231     | ILM      | 0.100:25   | 3 | None | GEO       |
| UMiami      | MCF7       | ILM      | 0.100:25   | 1 | None | GEO       |
| MAQC Site 1 | Mouse lung | ILM      | 100:75/25/0| 5 | Cubic| GEO       |
| MAQC Site 3 | Mouse lung | ILM      | 100:75/25/0| 5 | Cubic| GEO       |
| BIIB 500    | Mouse T cells | AFFX | 0.100:20   | 3 | MAS5 | GEO       |
| BIIB 100    | Mouse B cells | AFFX | 0.100:20   | 1 | MAS5 | GEO       |

Source: data source; type, tissue/cell types; platform, expression platform; proportion, $p_A$; n-number of samples at each proportion; Norm, normalization; GEO, GEO accession number. See text for further details.
Table 2. Bootstrap estimates of $p_A$

| Source         | Norm    | nb | Prop | Est  | SE    | 90% CI          |
|----------------|---------|----|------|------|-------|-----------------|
| UM-MDA231      | Q spline| 39 | 0.75 | 0.788| 0.023 | (0.746, 0.819)  |
|                |         | 35 | 0.50 | 0.529| 0.065 | (0.396, 0.604)  |
|                |         | 10 | 0.25 | 0.304| 0.108 | (0.180, 0.437)  |
| UM-MCF7        | Quantile| 100| 0.50 | 0.448| 0.057 | (0.375, 0.553)  |
|                |         |    |      |      |       | (0.286, 0.336)  |
| MAQC-ILM       | Cubic   | 40 | 0.75 | 0.776| 0.041 | (0.710, 0.842)  |
|                |         | 55 | 0.25 | 0.303| 0.021 | (0.275, 0.335)  |
| MAQC-AFFX      | MAS5    | 68 | 0.25 | 0.270| 0.027 | (0.232, 0.317)  |
|                |         |    |      |      |       | (0.688, 0.805)  |
| BiBiB00        | MAS5    | 100| 0.40 | 0.493| 0.053 | (0.388, 0.567)  |
|                |         |    |      |      |       | (0.80, 0.860)   |
| BiBiB100       | MAS5    | 100| 0.60 | 0.518| 0.050 | (0.437, 0.593)  |
|                |         |    |      |      |       | (0.722, 0.789)  |

Source, data source; Norm, normalization, nb, number of bootstrap samples; Prop, true value of $p_A$; Est, bootstrap point estimate; SE, bootstrap standard error; 90% CI, 90% bootstrap confidence intervals. Bold values denotes cases where the true $p_A$ is not in the interval.

tells us that a true 90% confidence interval is larger than the stated 90% confidence interval. In light of this the accuracy of our method is most likely better than the results in Table 2 would suggest.

An accurate estimate of $p_A$ can be used to generate estimates of expression specific to each tissue/cell type. Given expression from a mixed sample $AB$ and an estimate of $p_B$, we can estimate $E(A)$ and $E(B)$ as

$$E(A) = p_A E(AB) \quad \text{and} \quad E(B) = (1 - p_A) E(AB).$$

As we observe $E(A)$, we can compare $E(A)$ with the observed $E(A)$ to assess the quality of our estimate. Whether the error in using $E(A)$ as an estimate of $E(A)$ can be used to improve our estimate of $p_A$ is a topic for future research.

5 DISCUSSION AND CONCLUSIONS

We have demonstrated a statistical method for estimating the proportions of each sample (Samples A and B) in a two-sample mixture (AB). This method requires expression data generated from the mixed sample $AB$ and expression data generated from a purified sample of one type A. Given this information, the method approximates the proportion $p_A$ as the minimum of the ratios of expression in the mixed and purified samples, where the minimum is taken over genes. For this estimate to be accurate, it is required that the data be transformed; the value of the parameter of the transformation is determined by a geometric argument involving the minimum radius of curvature of a function, parameterized as a curve in the plane. Our results show that our method provides a reasonably accurate estimate of $p_A$ on both proprietary and publicly available datasets.

As demonstrated in Cleator et al. (2006) a large value of $p_A$ (say, over 0.5) can have a substantial effect on the results of tests for differential expression and confound tumor classification. However, whether a large $p_A$ should be cause for concern depends on the specific study. We would argue that $p_A$ should be assessed in all samples, but the action of the investigator in response to a large value of $p_A$ may vary from no action to discarding the sample from further consideration. In the case where $p_A$ is very large, our method will still give a reasonable estimate of $E(B) / E(AB)$ but the variability in this estimate could be large. Whether a large $p_A$ necessitates a renormalization of the data is unknown; we conjecture that if $E(A)$ and $E(A)$ are comparable then renormalization is unnecessary.

The results presented are preliminary and as such further research is required to optimize and validate our method. Our bootstrap point estimates and confidence intervals could be substantially improved by increasing the number of bootstrap samples $T$ and running diagnostics to ensure that the number of samples and size of samples are adequate for generating valid bootstrap quantities of interest (Canty et al., 2002). In addition, we would like to explore the relationship between the method of normalization and our estimation technique. By altering the noise distribution, normalization alters the relationship between the noise and the values of $R_i$, thereby influencing the accuracy of $\text{min} R_i$ as an estimate of $p_A$. The extent of this influence is unknown, but further research may help determine which normalization method yields the most accurate estimate of $p_A$. Finally, the calculation of the radius of curvature depends on the estimation of the second derivative of the curve; we approximate the second derivative by the second difference equation [Equation (4)]. This approximation is accurate if the curve is smooth and is well sampled, i.e. the distance between $x_k$ and $x_{k-1}$ is small. Using a well-sampled curve in our method can be computationally expensive if the range of values of $u$ (i.e. values of $s$) is large. We would like to design a variation of our method which starts with a sparsely sampled curve over a large range of values of $u$ and iteratively narrows the range of interest and increases the sampling density as information about the probable location of $s^*$ is obtained. This should yield a better estimate of $p_A$ at lower computational expense. We hope to implement this variation and provide our approach to the statistics community as an R package (R Development Core Team, 2009).

Our definition of the ‘elbow’ of a curve as the point of minimum radius of curvature is applicable to other problems in statistics, such as the choice of the number of principal components in a principal components analysis (Jolliffe, 2002). One existing way to make this choice is to identify the ‘elbow’ of the curve from a scatter plot and choose the number of components closest to the ‘elbow’. Our procedure for finding the minimum radius of curvature, coupled with a curve-fitting method, may be directly applicable to this problem. This would provide a formalization, in the spirit of Zhu and Ghodsi (2006), of what is currently an ad hoc approach.

ACKNOWLEDGEMENTS

We would like to thank the Center for Computational Science and the lab of Marc E. Lipman, MD, for their suggestions and input.

**Funding:** National Cancer Institute of the National Institutes of Health (K25 CA111636 to J.C.).

Conflict of Interest: none declared.

REFERENCES

Affymetrix Inc. (2009) Affymetrix Expression Console Software Version 1.0 — User Guide. Santa Clara, CA.
Ames,W. (1977) Numerical Methods for Partial Differential Equations. Academic Press, New York.
Barrett,T. et al. (2006) NCBI GED: mining tens of millions of expression profiles—database and tools. Nucleic Acids Res., 34, D760–D765.
Carly,A. (2002) Bootstrap diagnostics and remedies. Can. J. Stat., 30, 5–27.
Cauter,H. et al. (2005) Microarray gene expression data analysis: A beginner’s guide. Blackwell Science, Malden, MA.
Cleaton,S. et al. (2006) The effect of the stromal component of breast tumours on prediction of clinical outcome using gene expression microarray analysis. Breast Cancer Res., 8, B32.
Dunning,M. et al. (2008a) beadarray: R classes and methods for illumina bead-based data. Bioinformatics, 23, 2183–2184.
Dunning,M. et al. (2008b) Statistical issues in the analysis of illumina data. BMC Bioinformatics, 9, 85.
Elston,R. (1979) Bootstrap methods: Another look at the jackknife. Ann. Stat., 7, 1–26.
Fend,F. and Raffeld,M (2000) Laser capture microdissection in pathology. In Laser Capture Microdissection in Pathology, 1–26. San Diego, CA.
Fend,F. and Raffeld,M (2007) Electronically subtracting expression patterns from a mixed cell population. Statistical expression deconvolution. Bioinformatics, 23, 3328–3334.
Illumina Inc. (2009) beadStudio Gene Expression Module v3.4 User Guide (1117265 Rev A), San Diego, CA.
Johnstone,D. et al. (2008) Effects of different normalisation and analysis procedures on illumina gene expression microarray data. In Proceedings of the 19th International Conference on Genome Informatics (GIW 2008), Poster.
Jolliffe,I. (2002) Principal components analysis. 2nd edn. Springer, Berlin.
Koltsi,H. and Weingarten-Baren,C. (2008) Specificity of DNA microarray hybridization: characterization, effectors, and approaches for data correction. Nucleic Acids Res., 36, 2395–2405.
Lakshmanaki,H. et al. (2005) In silico microdissection of microarray data from heterogeneous cell populations. BMC Bioinformatics, 6, 54.
Leeb,H. (2009) Conditional predictive inference post model selection. Ann. Stat., 37, 2838–2876.
Lipschutz,M. (1969) Schaum’s outline of theory and problems of differential geometry. McGraw Hill, New York.
Lu,P. et al. (2003) Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations. Proc. Nat. Acad. Sci. USA, 100, 10370–10375.
MAQC Consortium (2006) The microarray quality control (MAQC) project shows inter- and inplatform reproducibility of gene expression measurements. Nat. Biotechnol., 24, 1151–1161.
Michiels,S. et al. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet, 365, 486–492.
Pittman,J. et al. (2004) Clinico-genomic models for personalized prediction of disease outcomes. Proc. Nat. Acad. Sci. USA, 101, 8431–8436.
R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
Shearin,J. et al. (2006) Accurate and precise transcriptional profiles from 50 pg of total RNA or 100 flow-sorted primary lymphocytes. Genomics, 88, 111–121.
Shen,X. et al. (2004) Inference after model selection. J. Am. Stat. Assoc., 99, 751–761.
Stuart,R. et al. (2004) In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. Proc. Nat. Acad. Sci. USA, 101, 615–620.
vant Veer,L. et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415, 530–536.
Vend,D. et al. (2001) Separation of samples into their constituents using gene expression data. Bioinformatics, 17, 5279–5287.
Wang,M. et al. (2006) Computational expression deconvolution in a complex mammalian organ. BMC Bioinformatics, 7, 328.
Wright,L. et al. (2008) The incredible shrinking world of DNA microarrays. Mol. Biosyst., 4, 726–732.
Zhu,M. and Ghosh,A. (2006) Automatic dimensionality selection from the sceer plot via the use of profile likelihood. Comput. Stat. Data Anal., 51, 818–830.