

Gene expression

Electronically subtracting expression patterns from a mixed cell population

Mark M. Gosink^{1,*}, Howard T. Petrie² and Nicholas F. Tsinoremas³¹Scientific Computing, ²Cancer Biology, Scripps Florida, 5353 Parkside Dr Jupiter, FL 33458 and³Center for Computational Science, University of Miami, Miller School of Medicine, Clinical Research Building, Suite 1188, 1120 NW 14th St., Miami, FL 33136

Received on July 20, 2007; revised on September 19, 2007; accepted on October 4, 2007

Advance Access publication October 22, 2007

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Biological samples frequently contain multiple cell-types that each can play a crucial role in the development and/or regulation of adjacent cells or tissues. The search for biomarkers, or expression patterns of, one cell-type in those samples can be a complex and time-consuming process. Ordinarily, extensive laboratory bench work must be performed to separate the mixed cell population into its subcomponents, such that each can be accurately characterized.

Results: We have developed a methodology to electronically subtract gene expression in one or more components of a mixed cell population from a mixture, to reveal the expression patterns of other minor or difficult to isolate components. Examination of simulated data indicates that this procedure can reliably determine the expression patterns in cell-types that contribute as little as 5% of the total expression in a mixed cell population. We re-analyzed microarray expression data from the viral infection of macrophages and from the T-cells of wild type and Foxp3 deletion mice. Using our subtraction methodology, we were able to substantially improve the identification of genes involved in processes of subcomponent portions of these samples.

Contact: gosink@scripps.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Expression profiling is a well-established technique for identifying global expression patterns within cells, and is used for purposes ranging from the identification of disease biomarkers to basic understanding of cellular processes (Tavazoie *et al.*, 1999; Yewdell and Bennink, 1999). Unfortunately, many biological samples contain mixtures of cell-types. For example, viruses infect only a proportion of the cells in a tissue, organs contain numerous cell-types (Ushiki, 1986; Woods and Walker, 1996), and cancer cells make up only part of a biopsy sample. Cleator *et al.* recently demonstrated that the non-cancerous portion of breast cancer samples can significantly

affect expression profiles, and that factoring in the amount of cancerous material in samples can improve the accuracy of response prediction (Cleator *et al.*, 2006). In all these examples, the cell-type of interest (infected cells, cancer cells or specific components of an organ) is only a subset of cells in the sample. This severely limits the conclusions that can be made about the specificity of gene expression in the cell-type of interest.

Techniques such as laser capture microdissection (LCM) allow isolation of regions of a biological sample that are separated by as little as a few cell widths. Such samples can then be analyzed by expression analysis. However, LCM requires the cell-type of interest to be morphologically distinguishable and physically separated from the other cell-types in the sample. Finally, it can be very time consuming and requires specialized equipment to obtain a sufficient quantity of some cell-types to perform expression profiling. RNA amplification procedures can be used but these introduce artifacts of amplification (Mills *et al.*, 2001). Cell sorting can also be used to isolate cells of interest; however, this technique requires a suitable biomarker for the cell-type to have been previously identified. In addition, both of these techniques are limited in that LCM is only practical with solid tissues, while flow sorting is only applicable for cells that can be put into suspension. Finally, the act of separation itself can result in the alteration of expression patterns.

In this article, we describe an algorithmic approach to calculate the expression profile of a cell-type of interest from that of a mixed cell population consisting of at least two types of cells. This technique electronically subtracts the expression profile of one component of a sample from the expression profile of the total sample, and thus revealing the profile of the second component. We demonstrate the utility of this approach using simulated expression data. We further apply this process to publicly available experimental datasets found in Gene Expression Omnibus (GEO) (Barrett *et al.*, 2005) to further validate this process. More specifically, we describe the subtraction of the expression profile of an uninfected sample from the profile of a mixed virally infected/uninfected sample to yield the expression profile of the infected cells alone. We also describe the subtraction of the expression profile of T-cells from a Foxp3 deletion mouse from those of a wild-type animal to yield the expression of regulatory T-cells (T-reg). (Foxp3 gene is essential for the development of a subset of T-cells known as T-reg cells.)

*To whom correspondence should be addressed.

2 METHODS

2.1 Simulated expression data

Simulated expression data was based on expression profiles downloaded from the GEO public repository at the National Center for Biotechnology Information (NCBI) (Barrett *et al.*, 2005). Data from a number of pairs of expression profiles from the GEO database were used as theoretical pure cell-type samples. The pair combinations, seed arrays, used were: GSM17128/GSM17214; GSM18915/GSM18939; GSM18921/GSM18917; GSM18927/GSM18905; GSM18931/GSM19019; GSM18933/GSM18925; GSM18933/GSM18977; GSM18963/GSM18965; GSM18977/GSM18979 and GSM18977/GSM18981. Intensity values from these pairs were used as the ‘true’ expression values of a known and unknown sample. A ‘mixed’ cell population expression profile was generated from each pair using Equation (5) at proportion (pA) values of: 0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 0.975 and 0.99. From each of these true expression profiles [known (A), unknown (B) and mixed (AB)], three noisy samples were generated using a noise model we built from experimental data. Briefly, for a given intensity I , an error is generated from:

$$\text{error} = (0.11294 \times (I^{0.9864})) \times \text{ND}$$

where ND is a randomly generated normal distribution factor with a mean of 0.0 and a SD of 1.5. The error is then added to the intensity value to yield a noisy intensity. The error model does a good job of generating simulated noisy samples over the range of intensities normally seen with the Affymetrix arrays described here. A comparison of expression profiles between multiple samples from real samples versus those generated by the above method could not be distinguished from each other (http://compsci.florida.scripps.edu/Elec_Sub_Paper/Simulated_Data.tar.gz).

2.2 Non-simulated data

The T-reg cell was extracted from the GEO dataset—GDS1113. Only samples GSM44982 and GSM44980 were used in this analysis. Expression data for virally infected cells were extracted from GEO dataset GDS1271.

2.3 Determination of proportion, pA

If a set of genes can be identified which are known to express exclusively in one cell-type in a mixed sample, the expression of each of these genes in the known sample (A) and the mixed sample (AB) can be used to calculate a value for pA using:

$$pA = E_{iAB}/E_{iA} \quad (1)$$

However, in cases where a suitable set of genes that express exclusively in the known cell-type cannot be identified or as additional confirmation, the pA of expression in the mixed sample (AB) can be calculated using the following methodology. A mixed to pure known ratio, $R_{\text{mix/pure}}$, for all genes is calculated from:

$$R_{\text{mix/pure}} = E_{iAB}/E_{iA} \quad (2)$$

Genes that express at low levels in the known sample (below the chip mean intensity) are eliminated. All the remaining values are ranked from low to high. The ratios values at the lowest end of this ranked list are proportional to the actual value of pA. For these analyses with Affymetrix chip data, from the smallest value we used the ratio value at 5% in the following equation to calculate the proportion:

$$pA = -0.01831 + 0.3485 \times R_{\text{mix/pure}5\%} + 1.182 \times (R_{\text{mix/pure}5\%})^2 \quad (3)$$

2.4 Calculating expression in the unknown sample

The expression intensity for any one gene in mixture of two cell-types (A + B) can be defined using a mixture model where the intensity is equal to the sum of the intensities of that gene from each cell-type. In other words, the expression of any gene i in the mixture can be calculated from the equation:

$$E_{iAB} = pA \times E_{iA} + (1 - pA) \times E_{iB} \quad (4)$$

where E_{iAB} is the expression of that gene as measured in the mixed sample, E_{iA} is the expression of gene i in cell-type A and E_{iB} is the expression of gene i in cell-type B. The proportion of expression in the mix due to cell-type A is given by pA. Once a value for pA was established, the intensity value for each gene (probeset) i in the sample, the expression is calculated using a rearrangement of Equation (4):

$$E_{iB} = (E_{iAB} - (pA \times E_{iA})) / (1 - pA) \quad (5)$$

2.5 Determining over-representation of Gene Ontology categories in gene lists

We analyzed the various gene lists using Fisher’s exact test as implemented in the R statistics package to identify Gene Ontology (GO) categories that were over-represented in each list (Team, 2004). We corrected for multiple testing using Storey’s q -value function from within R to generate q -values (Storey and Tibshirani, 2003).

2.6 Determining of the area under the ROC curves for gene expression

True positive rates and false positive rates were calculated as described by Fawcett (Fawcett, 2006). True positives were defined as genes whose expression was 2-fold or more over-expressed in the unknown sample versus the known sample. Calculated fold-change values were generated for each gene from the calculated expression in the unknown sample over the expression in the known sample. The area under the receiver operating characteristic (ROC) curve was estimated by integrating over the calculated fold-change values. The area under the ROC was calculated for each seed set pair and over pA values from 0.00 to 0.95.

3 ALGORITHM

3.1 Development of the electronic-subtraction methods and testing using simulated expression data

If, in a two-component mixture, the expression profile of one component and the proportion of expression it contributes to the mixture are known, then it is relatively straight forward to subtract the expression of the known sample’s expression from the mixture’s expression to yield the unknown component’s expression profile. Unfortunately, for most samples, the portion of expression coming from each cell-type is also unknown making the calculation of the expression pattern of the second cell-type impossible to determine. In a few cases, a set of genes may be known to express exclusively in one of the cell-types. These genes may then be used to estimate the proportion of expression coming from that component. This approach was also reported by Lu *et al.* with their work on yeast-cell profiling (Lu *et al.*, 2003). Gosh describes a mixture model method utilizing a pathologist’s assessment of the percent cancer in a sample to establish the proportion of expression in a sample coming from the cancer cells

(Ghosh, 2004). Wang *et al.* and Lahdesmaki *et al.* both use expression data from purified reference cell-types to determine the proportion of each cell-type in heterologous samples (Lahdesmaki *et al.*, 2005; Wang *et al.*, 2006). We initially used a set of genes known to be exclusive to one cell-type but during this analysis, we observed that this pA, could also be approximated from the lowest ratio value of the intensity from the mixed sample genes divided by the intensity from the known sample genes. We explored this relationship using a number of simulated expression datasets. Further analysis with simulated noiseless expression data indicated that the minimal ratio value always approached the pA value. For noiseless expression data, the equivalency of the minimal ratio value and pA can be demonstrated mathematically.

A mixed to pure ratio, $R_{\text{mix/pure}}$, is calculated for each gene i from:

$$R_{\text{mix/pure}} = E_{iAB}/E_{iA} \quad (6)$$

By substituting in the values from Equations (5) and (6) can be re-arranged to

$$R_{\text{mix/pure}} = (pA \times E_{iA} + (1 - pA) \times E_{iB})/E_{iA} \quad (7)$$

or

$$R_{\text{mix/pure}} = (pA \times E_{iA})/E_{iA} + ((1 - pA) \times E_{iB})/E_{iA} \quad (8)$$

For genes expressing weakly in the known cell-type (A) and strongly in the unknown cell-type (B) the first term of Equation (8), ' $pA \times E_{iA}/E_{iA}$ ', becomes negligible. The second term, ' $(1-pA) \times E_{iB}/E_{iA}$ ', tends towards infinity as E_{iB} becomes large and E_{iA} becomes small (i.e. for cell-type B specific genes). Genes expressing equally in both tissues have $R_{\text{mix/pure}}$ values tending towards 1. For genes with significant expression in cell-type A and little or no expression in cell-type B ($E_{iB} = 0$), Equation (8) simplifies to:

$$R_{\text{mix/pure}} = (pA \times E_{iA})/E_{iA} + 0/E_{iA} \quad (9)$$

or

$$R_{\text{mix/pure}} = pA \times (E_{iA}/E_{iA}) \quad (10)$$

or

$$R_{\text{mix/pure}} = pA \quad (11)$$

Therefore, the minimum value for $R_{\text{mix/pure}}$, where E_{iA} is significant, can be assumed to be equivalent to pA. If artificial expression data is generated such that the expression profile in cell-type A is independent of the expression profile in cell-type B, this $R_{\text{mix/pure}}$ value approaches pA in a near linear fashion near the lowest ratios (data not shown). With real expression data and with our simulated expression data, there are few genes that are absolutely specific to one cell-type, so $R_{\text{mix/pure}}$ approaches pA in an inverted S-curve.

Figure 1 demonstrates the relationship between $R_{\text{mix/pure}}$ and pA using simulated data without added sample noise for one pair of seed experiments. Other seed pairs show a similar relationship (data not shown). In addition to the relatively scarcity of absolutely cell-type-specific genes, in non-simulated datasets the experimental noise results in some genes from either the mixed sample or the cell-type A sample that have aberrantly high or low values. Further error can be introduced if the expression in the cell-type B sample used and the

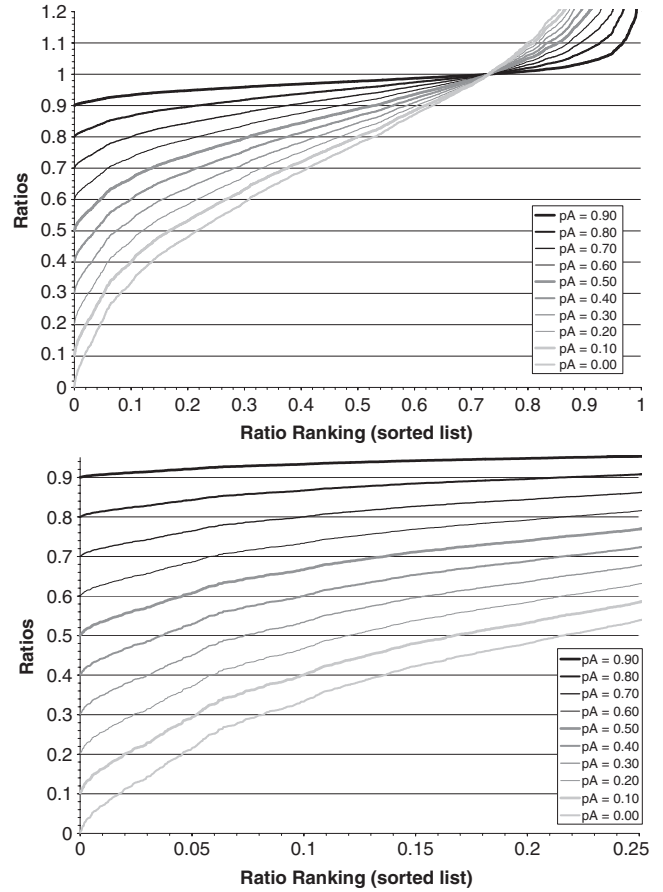


Fig. 1. Rank sorted ratios ($R_{\text{mix/pure}}$) for genes from noiseless simulated datasets at a range of proportion values, pA. Rank sorted ratios are generated using a single set of seed arrays (GSM18977/GSM18979). (Top) Rank sorted ratios at a range of proportion values, pA. (Bottom) Close-up view of y-intercept region.

expression in cell-type B in the mixed sample are different. For example, if the isolation of cell-type B results in a change of its expression pattern before it can be measured or if cell-type B originates from a different sample than the mixture further error will be introduced. These errors will result in a number of genes with aberrant $R_{\text{mix/pure}}$ values. To determine if the proportion of expression from cell-type A, pA, could reliably be calculated from noisy data, we analyzed a number of different combinations of expression pattern sets with simulated noisy data.

These simulated noisy expression datasets were generated from pairs of expression samples from NCBI's GEO database where one sample represented the known cell-type (A) and the other represented the unknown sample (B). These expression profiles were used as expression value seeds to generate a mixed (AB) expression profile (see Methods section for details). Noisy profiles were generated from the seeds' profiles and from the generated mixed cell-type profiles by adding simulated noise. Lists of sorted $R_{\text{mix/pure}}$ values for genes with cell-type A intensities greater than the chip median value, from smallest to highest, were generated for every probeset. Analysis of the

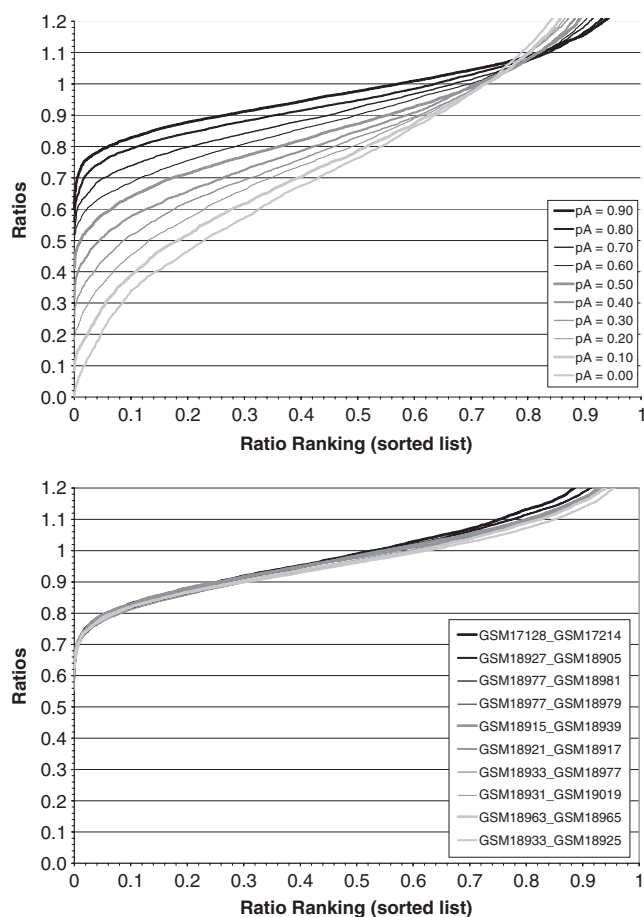


Fig. 2. Rank sorted ratios ($R_{\text{mix/pure}}$) for genes from simulated datasets with added noise. (Top) Sorted ratios using a single set of seed arrays (GSM18977/GSM18979) over a range of pA values. (Bottom) Sorted ratios using constant pA (0.9) using different seed experiment combinations.

$R_{\text{mix/pure}}$ values generated from simulated datasets at series of pA values is shown in Figure 2. The data indicate that there is a relationship between the ratio list and the pA of the ‘purifiable’ cell-type in the mixed sample. We also examined the effects of using different seed sets at a constant proportion. Figure 2 shows the variability in the ratio list using different seed arrays but at a constant proportion.

To develop an equation to calculate pA from the ratio data, ratio datasets were generated from 10 different combinations of seed arrays over a series of pA values ranging from 0.00 to 1.00. Ratio values were extracted from each sorted ratio dataset at a series of positions within each ranked list. The extracted values from a given position within the lists were plotted against the pA values used to generate the data. For example, the ratio values extracted at 5% index of the ranked list plotted against pA value are shown in Figures 2(top) and 3. Regression analysis was performed using a variety of algorithms and using values sampled at a range of positions within the list. We also evaluated the use of multivariate regression using multiple positions in the ranked list (data not shown). We evaluated the

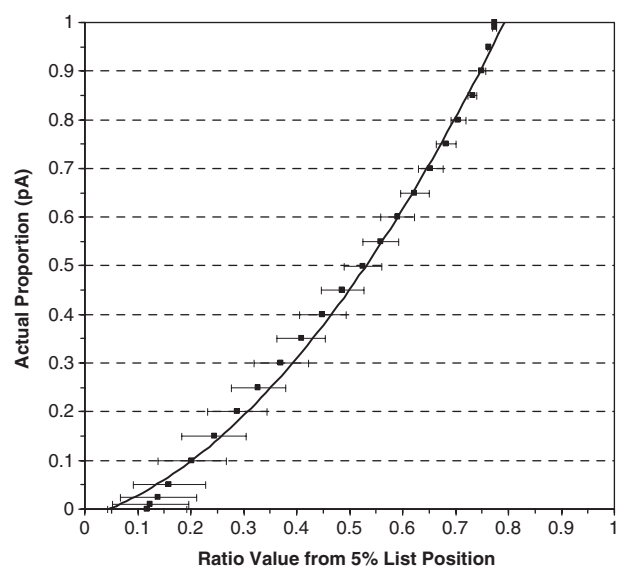


Fig. 3. Mean proportion values versus ratio values selected from the 5% (0.05) position in the rank sorted ratio lists. Expression datasets were generated from a 10-seed array combinations over a range of pA values.

r^2 error for each of the algorithms and for each ranked list position samples. For the microarray data used in our analyses, a second-order polynomial of the 5% value of the ranked lists yielded the best results with an r^2 of 0.977 [Equation (3)].

The calculated values correlated well over most values of pA with the most deviation at the extreme values. Since other types of data would have differing error models and differing numbers of cell-type specific genes, the regression algorithm and position within the ranked list should be re-evaluated.

4 RESULTS

4.1 Analysis of simulated data

Equation (3) was used to calculate pA values from simulated data generated from all expression seed array pairings over a range of actual pA values. The actual proportion was plotted against the mean calculated values and their SDs in Figure 4. Linear regression analysis performed on the calculated ratios versus the actual ratios resulted in an r^2 value for of 0.995 indicating that the electronic-subtraction method accurately calculates pA values. We also performed the reverse operation and calculated the proportion of expression from cell-type B. The calculated proportions for B had an inverse relationship to the calculated values for cell-type A (see Supplementary Fig. 1). Once the pA value was determined, the expression of each gene in the unknown sample could then be calculated from Equation (5).

A number of methodologies have been developed to identify differentially expressed genes when numerous replicates of each sample type exist. Some of these approaches require an estimate of the proportion of expression contributed from each cell-type. Our approach could provide a means to determine the proportional value. However, for many samples few

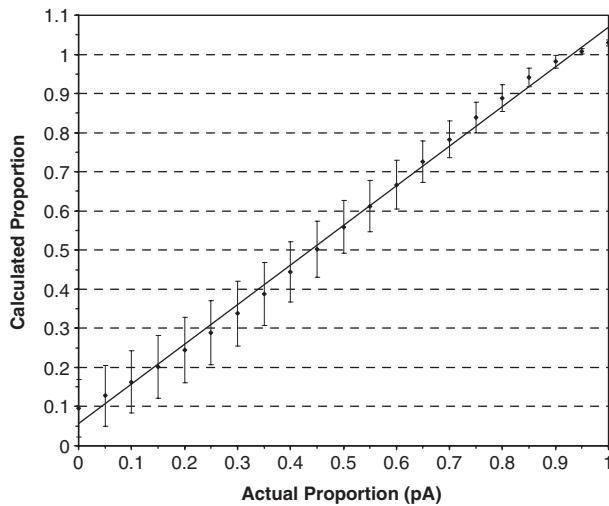


Fig. 4. A comparison of actual pA versus mean calculated values for all simulated expression datasets over a range of pA values.

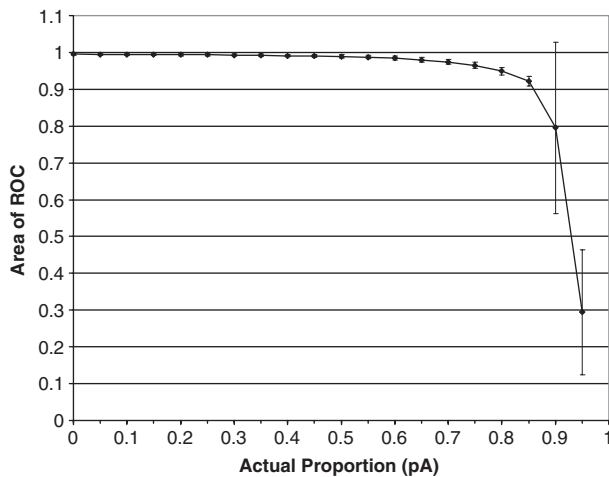


Fig. 5. Area under the curve for the ROC of the subtractive approach in the detection of genes over-expressing >2-fold in unknown samples.

replicates exist. In such cases, a 2-fold-change is often utilized as a cut-off for changed genes. A variety of methods have been employed to determine which genes are differentially expressed between two samples (Jeffery *et al.*, 2006). For this analysis we chose to use fold-change. Specifically, genes that were more than 2-fold elevated in the unknown sample as compared to the known sample over a range of mixture fractions were identified. We calculated the area under the curve from a ROC plot for pA values from 0.05 to 0.95 for each seed set pair. The average area of the ROC and their SDs are shown in Figure 5. The results indicate that the predictive ability of this method remains high until the proportion of the unknown sample falls below 15% of the total. The majority of uncertainty is due to the over-subtraction of some expression values resulting in genes with negative expression values. If calculated expression data with values below 1 are removed the area of the

ROC is 0.88 and 0.66 at pA values 0.90 and 0.95, respectively (see Supplementary Fig. 2).

4.2 Analysis of SeV-infected cell data

We re-examined two expression datasets from NCBI's GEO database to determine if we could identify additional relevant information. The datasets were chosen to represent different kinds of mixed samples commonly seen in experimental data.

The first set was from an experiment to determine the effects on macrophage survival upon infection with *Sendai* virus (SeV) by Tyner *et al.* (Tyner *et al.*, 2005). The samples consist of macrophages mock-infected with UV-irradiated SeV and a second of macrophages infected with viable SeV. As with most viral infections, the rate of infection was not 100%, so this sample contained both infected and uninfected cells (see Fig. 5b in the Supplementary Material for Tyner *et al.*).

We applied the electronic-subtracted expression profile of the uninfected macrophages from the partially infected sample to determine the expression profile of the infected macrophages alone. We used our method to identify the proportion of expression from the infected cells as 0.667. This fraction approximates the rate of infected cells as visualized by SeV immunostain versus DAPI (Tyner *et al.*, 2005). The electronic-subtraction method was used to calculate the expression patterns in the infected cells alone. We then identified two groups of genes specific to virally infected cells. The first group consisted of those genes identified by straight 2-fold or more expression in the SeV-infected macrophage cells versus the mock-infected cells. We set a minimal intensity value in the mock-infected sample as >11 (the median intensity value). Under these criteria, 124 probesets (113 genes) are identified as specific to virally infected macrophages. These elevated genes include the *Ccl5* gene, which was reported by the authors. The second group of identified genes were those meeting the same criteria but using the electronic-subtraction calculated values for the infected cells. Using the same criteria as above except the calculated values were used for the SeV-infected cells, 474 probesets (399 genes) are identified as specific to virally infected macrophages. Selected GO categories that are over-represented in the two gene lists are shown in Table 1 (see Supplementary Material for complete lists). In their paper, the authors demonstrate that the elevated *Ccl5* gene has a role in the regulation of apoptosis through the *Gαi*-PI3K-AKT and *Gαi*-MEK-ERK pathways. Both methodologies identify gene lists with a statistically significant over-representation of probesets belonging to the 'GO:0006955 immune response'. Our subtraction method identifies the GO categories 'GO:0006915 apoptosis' and 'GO:0042981 regulation of apoptosis' with *q*-values of $8.04E-04$ and $1.05E-02$, respectively. The standard $2\times$ elevated method does not yield statistically significant scores for these categories with *q*-values of $1.16E-01$ and $7.88E-02$, respectively. MHC I proteins are known to play a role in the destruction of virally infected cells and a number of viruses interfere with the expression of the proteins in this complex (Hewitt, 2003; Kunisawa *et al.*, 2001; Yewdell and Bennink, 1999). Using the subtraction method, five genes belonging to the GO category 'GO:0002474 antigen processing and presentation of peptide antigen via MHC class I' were

Table 1. Comparison of selected gene ontology categories over-represented in genes identified using either the unsubtracted values 'Infected Over-expressed' or the electronic-subtraction values 'Infected Over-expressed (calculated)' for over-expressed genes in SeV-infected macrophages

GO categories	Number of genes	<i>q</i> -Value
Infected over-expressed		
Immune response	25	1.49E-12
Antigen processing and presentation of peptide antigen	5	1.15E-2
Antigen processing and presentation of peptide antigen via MHC class I	2	5.98E-1
Apoptosis	11	1.16E-1
Regulation of apoptosis	9	7.88E-2
Calculated infected over-expressed		
Immune response	37	3.89E-8
Antigen processing and presentation of peptide antigen	8	1.05E-2
antigen processing and presentation of peptide antigen via MHC class I	5	8.41E-2
Apoptosis	33	8.04E-4
Regulation of apoptosis	22	1.05E-2

identified for a marginally significant *q*-value of 0.084. The standard method only identifies two genes in this category for a *q*-value of 0.598.

4.3 Analysis of regulatory T-cell data

We also re-examined expression data from T-cells, one sample was from wild-type T-cells and the second sample was from T-cells from a Foxp3 deletion mouse (Fontenot *et al.*, 2005). Wild-type T-cells contain a mixture of T-reg and non-T-reg cells. Foxp3 is required for the development of subset of T-cells known as 'T-reg' cells (Hori *et al.*, 2003; Wildin *et al.*, 2001). T-reg cells control the response of other T-cells and block self-recognition by these cells. Animals and people with a mutated FOXP3 gene do not make T-reg cells and succumb to autoimmune disease at an early age (Wildin *et al.*, 2001). Fontenot *et al.* isolated purified T-cells from wild type and Foxp3 knockout mice and performed microarray analysis on each (Fontenot *et al.*, 2005). We re-examined their data, subtracting the non-T-reg T-cells from the Foxp3 deletion animal from the mixed T-cells of the wild-type animal to identify genes, which were up regulated in T-reg cells. The proportion of expression due to the T-reg cells in the wild-type mouse was calculated to be 0.548 by our method.

We identified two groups of over-expressed genes. The first set was defined as those genes, which were expressing two-fold or more in the wild-type sample as compared to the FOXP3 K.O. sample. Low-expressers with intensities less than the median value (<6) in the FOXP3 K.O. sample were removed. Table 2 displays the number of genes identified and the calculated significance for a number of Gene Ontology categories thought to be involved in T-reg processes.

Table 2. Comparison of select Gene Ontology categories over-represented in genes identified using either the unsubtracted values 'T-reg Over-expressed' or the electronic-subtraction values 'T-reg Over-expressed (calculated)' for genes over-expressed in Regulatory T-cells

GO categories	Number of genes	<i>q</i> -Value
Regulatory T-cell over-expressed		
Immune system process	34	1.67E-1
Cytokine production	5	1.0
Protein transport	38	3.60E-1
Ras protein signal transduction	13	3.13E-2
Regulation of Ras protein signal transduction	8	7.36E-2
Calculated regulatory t-cell over-expressed		
Immune system process	89	1.50E-2
Cytokine production	20	2.33E-2
Protein transport	117	1.13E-4
Ras protein signal transduction	33	1.59E-4
Regulation of Ras protein signal transduction	17	1.13E-2

A complete list of all GO categories is available in the Supplementary Material. The number of probesets in group 2 (3914 probesets; ~2767 genes), was dramatically larger than that identified in group 1 (1229 probesets, ~939 genes). Table 2 reveals that a number of GO categories are over-represented in both groups. For example, the authors or the microarray data as well as others have reported that cytokines play a critical role in the function of T-reg cells (Wan and Flavell, 2006). We see evidence of cytokine involvement with the significant over-representation of genes belonging to the GO category 'cytokine production' using the subtraction method but not with the standard fold-change. Other groups have indicated that one role of T-reg cells is to control other immune cells through the regulated endocytosis of granzyme b (Gondek *et al.*, 2005; Sugimoto *et al.*, 2006). Additionally, Mead *et al.* demonstrated that the regulated exocytosis of CTLA-4 is dependent on the ras-family protein ARF-1 (Mead *et al.*, 2005). We see evidence of this with the over-representation of the GO categories 'ras protein signal transduction' and 'regulation of Ras protein signal transduction'. Both categories are significantly over-represented using the subtractive method but only the first is identified as significant by the standard method.

5 DISCUSSION

Biological samples from higher organisms frequently contain a number of different cell-types. Each cell-type may play a distinct role in the function of that sample and have a distinct pattern of gene expression. If one is interested in the function of one of these underlying cell-types, it is critical that the cells of interest be isolatable. Unfortunately, this can be a complex process and it is not always possible to physically isolate every cell-type within a sample. If the proportions of the sample's expression profile that each cell-type contributes to that profile can be determined, the expression profiles of the individual components can be determined. We have developed a method

to determine the proportion of expression of an underlying cell-type within a sample without advance knowledge of the expression of any genes within that sample. We demonstrate the subtraction methodology with the removal of a single isolatable component of a mixture with microarray data, but this technique could be expanded to the subtraction of multiple isolatable components. Proportion values could be estimated for known samples by sequentially generating ranked lists of ratio values (see Supplementary Material 1).

Our method was developed using expression data where few samples are available, we are currently working to refine this approach to yield better understanding of error in situations where multiple samples are available. The approach has potential applications in basic research as well as biomarker discovery. For example, the method could be applied to biopsy samples to remove the normal tissue profile to yield that of the cancerous cell profile. Finally, this general method should be applicable to other expression technologies such as proteomics profiling.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Sandra Cervino for advice on statistical analysis. They also wish to thank NCBI and the authors of the viral and regulatory T-cell datasets for making this information publicly available. M.M.G. and N.F.T. are supported by the Florida Funding Corporation. H.T.P. is supported by PHS grants AI/AG33940, AI67453 and AI/HL64665.

Conflict of Interest: none declared.

REFERENCES

- Barrett, T. *et al.* (2005) NCBI GEO: mining millions of expression profiles – database and tools. *Nucleic Acids Res.*, **33**, D562–566.
- Cleator, S.J. *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**, R32.
- Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874.
- Fontenot, J.D. *et al.* (2005) Regulatory T cell lineage specification by the forkhead transcription factor foxp3. *Immunity*, **22**, 329–341.
- Ghosh, D. (2004) Mixture models for assessing differential expression in complex tissues using microarray data. *Bioinformatics*, **20**, 1663–1669.
- Gondek, D.C. *et al.* (2005) Cutting edge: contact-mediated suppression by CD4+CD25+ regulatory cells involves a granzyme B-dependent, perforin-independent mechanism. *J. Immunol.*, **174**, 1783–1786.
- Hewitt, E.W. (2003) The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology*, **110**, 163–169.
- Hori, S. *et al.* (2003) Control of regulatory T cell development by the transcription factor Foxp3. *Science*, **299**, 1057–1061.
- Jeffery, I.B. *et al.* (2006) Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, **7**, 359.
- Kunisawa, J. *et al.* (2001) Sendai virus fusion protein mediates simultaneous induction of MHC class I/II-dependent mucosal and systemic immune responses via the nasopharyngeal-associated lymphoreticular tissue immune system. *J. Immunol.*, **167**, 1406–1412.
- Lahdesmaki, H. *et al.* (2005) In silico microdissection of microarray data from heterogeneous cell populations. *BMC Bioinformatics*, **6**, 54.
- Lu, P. *et al.* (2003) Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc. Natl Acad. Sci. USA*, **100**, 10370–10375.
- Mead, K.I. *et al.* (2005) Exocytosis of CTLA-4 is dependent on phospholipase D and ADP ribosylation factor-1 and stimulated during activation of regulatory T cells. *J. Immunol.*, **174**, 4803–4811.
- Mills, J.C. *et al.* (2001) DNA microarrays and beyond: completing the journey from tissue to cell. *Nat. Cell Biol.*, **3**, E175–178.
- Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
- Sugimoto, N. *et al.* (2006) Foxp3- dependent and -independent molecules specific for CD25+CD4+ natural regulatory T cells revealed by DNA microarray analysis. *Int. Immunol.*, **18**, 1197–1209.
- Tavazoie, S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.
- Team, R.D.C. (2004) R: a language and environment for statistical computing. R Foundation for Statistical Computing. *Bioinformatics*, **23**, 641–643.
- Tyner, J.W. *et al.* (2005) CCL5-CCR5 interaction provides antiapoptotic signals for macrophage survival during viral infection. *Nat. Med.*, **11**, 1180–1187.
- Ushiki, T. (1986) A scanning electron-microscopic study of the rat thymus with special reference to cell types and migration of lymphocytes into the general circulation. *Cell Tissue Res.*, **244**, 285–298.
- Wan, Y.Y. and Flavell, R.A. (2006) The roles for cytokines in the generation and maintenance of regulatory T cells. *Immunol. Rev.*, **212**, 114–130.
- Wang, M. *et al.* (2006) Computational expression deconvolution in a complex mammalian organ. *BMC Bioinformatics*, **7**, 328.
- Wildin, R.S. *et al.* (2001) X-linked neonatal diabetes mellitus, enteropathy and endocrinopathy syndrome is the human equivalent of mouse scurfy. *Nat. Genet.*, **27**, 18–20.
- Woods, G.L. and Walker, D.H. (1996) Detection of infection or infectious agents by use of cytologic and histologic stains. *Clin. Microbiol. Rev.*, **9**, 382–404.
- Yewdell, J.W. and Bennink, J.R. (1999) Mechanisms of viral interference with MHC class I antigen processing and presentation. *Annu. Rev. Cell Dev. Biol.*, **15**, 579–606.