BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 5 2014, pages 682–689
doi:10.1093/bioinformatics/btt566

Gene expression

Advance Access publication October 1, 2013

MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples

David A. Liebner[1,2,*], Kun Huang[2,3] and Jeffrey D. Parvin[2,3]

[1]Division of Medical Oncology, Department of Internal Medicine, [2]Department of Biomedical Informatics and [3]Comprehensive Cancer Center, Biomedical Informatics Shared Resource, The Ohio State University, Columbus OH 43210, USA

Associate Editor: Janet Kelso

ABSTRACT

Background: One of the significant obstacles in the development of clinically relevant microarray-derived biomarkers and classifiers is tissue heterogeneity. Physical cell separation techniques, such as cell sorting and laser-capture microdissection, can enrich samples for cell types of interest, but are costly, labor intensive and can limit investigation of important interactions between different cell types.

Results: We developed a new computational approach, called microarray microdissection with analysis of differences (MMAD), which performs microdissection *in silico*. Notably, MMAD (i) allows for simultaneous estimation of cell fractions and gene expression profiles of contributing cell types, (ii) adjusts for microarray normalization bias, (iii) uses the corrected Akaike information criterion during model optimization to minimize overfitting and (iv) provides mechanisms for comparing gene expression and cell fractions between samples in different classes. Computational microdissection of simulated and experimental tissue mixture datasets showed tight correlations between predicted and measured gene expression of pure tissues as well as tight correlations between reported and estimated cell fraction for each of the individual cell types. In simulation studies, MMAD showed superior ability to detect differentially expressed genes in mixed tissue samples when compared with standard metrics, including both significance analysis of microarrays and cell type-specific significance analysis of microarrays.

Conclusions: We have developed a new computational tool called MMAD, which is capable of performing robust tissue microdissection *in silico*, and which can improve the detection of differentially expressed genes. MMAD software as implemented in MATLAB is publically available for download at http://sourceforge.net/projects/mmad/.

Contact: david.liebner@gmail.com

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

Received on January 2, 2013; revised on August 15, 2013; accepted on September 25, 2013

1 INTRODUCTION

Gene expression microarrays have yielded valuable biological and medical insight into many disease processes and have been heavily researched, especially in the investigation of cancer biology (Alizadeh *et al*., 2000; Beer *et al*., 2002; Belbin *et al*., 2002; Bittner *et al*., 2000; Dhanasekaran *et al*., 2001; Meyniel *et al*., 2010; Perou *et al*., 2000; Sorlie *et al*., 2001; Stratford *et al*., 2010; van 't Veer *et al*., 2002). However, standard approaches to microarray analysis can be easily confounded by cellular heterogeneity in tissue samples and by variability in cell type composition. This may introduce false-positive correlations, but more importantly may also mask changes in gene expression within a given cell type of interest. Of particular concern is impact of sample heterogeneity on developing and validating microarray-based predictive and prognostic models for human diseases (Cleator *et al*., 2006; Debey *et al*., 2004; Elloumi *et al*., 2011; Feezor *et al*., 2004). Physical cell separation methods such as cell sorting and/or laser-capture microdissection can enrich samples for the cell type of interest, but such methods are time and resource intensive and can affect the quantity and quality of RNA available for subsequent analysis (Debey *et al*., 2004; Venet *et al*., 2001).

Computational microdissection using various statistical approaches has consequently been a subject of interest for several groups. Most techniques extend on the linear model of Venet *et al*. (2001), which estimates the final measured gene expression as the sum of gene expression of the contributing cell types. Several approaches estimate relative fractions of individual cell types within a sample using gene expression profiles that are characteristic for each cell type (Abbas *et al*., 2009; Ahn *et al*., 2013; Bolen *et al*., 2011; Gaujoux and Seoighe, 2012; Gong *et al*., 2011; Lu *et al*., 2003; Wang *et al*., 2006; Zhong *et al*., 2013), whereas other models estimate the characteristic gene expression profiles of each cell type using measured cell type fractions (Shen-Orr *et al*., 2010; Stuart *et al*., 2004). Simultaneous estimates of cell type expression profiles and cell type fraction, analogous to principal-components analysis have also been proposed (Erkkila *et al*., 2010; Lahdesmaki *et al*., 2005; Repsilber *et al*., 2010; Venet *et al*., 2001). A software package that incorporates several of these methodologies in a unified interface has recently been made available (Gaujoux and Seoighe, 2013).

We developed a flexible new model, microarray microdissection with analysis of differences (MMAD), which incorporates several features of the previously described approaches in addition to several novel features, designed to improve performance and utility. These include adjustments for reporting bias and

*To whom correspondence should be addressed.

normalization bias, incorporation of information theory criteria to minimize overfitting and methods for comparing gene expression between classes, either globally or on a cell-specific basis. An overview of MMAD is provided in Supplementary Figure S1.

2 METHODS

2.1 Deconvolution model

Let X_{ij} be the observed microarray expression value for gene (or probe) i and sample j, and x_{ij} be the $\log_2$-transform of X_{ij}. Let I denote the number of genes (or probes) and J denote the number of samples. If we assume that there are K cell types that exist in varying mixing proportions in the given samples, then we can model the observed gene expression profile as a linear combination of contributing cell types with a log-normally distributed error term:

$$x_{ij} = \log_2\left(\sum_{k=1}^{K} C_{ik} f_{kj}\right) + e_{ij} \tag{1}$$

Here, C_{ik} is the characteristic gene expression for gene i with respect to cell type k, and f_{kj} is the fraction of total messenger RNA (mRNA) contributed by cell type k in sample j. Of note, although we refer to C and f as characterizing the gene expression profiles of individual 'cell types', they can more generally characterize any source of mRNA with a characteristic expression profile that is present in varying concentrations in a set of samples. These could include individual cell types (e.g. fibroblasts, cardiomyocytes, adipocytes), different cellular states (e.g. cells entering mitosis, cells in G0) or even pure sources of mRNA mixed within a laboratory setting. For simplicity, we will continue to refer to these as 'cell types' for the remainder of this article.

2.2 Adjustment for bias in reported cell fraction

Unfortunately, there is no guarantee that the estimated cell fraction provided by the investigator corresponds directly to the measurable fraction of RNA contributed by each cell type. We consider the (not infrequent) case in which the cellular fraction (as assessed by cell count) is used as a surrogate for the RNA fraction, f. Let N_{kj} represent the number of cells of type k in sample j. If the measurable RNA content per cell, ρ_k, varies for each cell type k, the reported value of f will be biased. That is,

$$f_{kj,\ \text{reported}} = \frac{N_{kj}}{\sum_q N_{qj}}, \text{ whereas } f_{kj,\ \text{actual}} = \frac{\rho_k N_{kj}}{\sum_q \rho_q N_{qj}} \tag{2}$$

To account for this, we introduce the concept of the effective RNA fraction f', which replaces f in Equation (1), and which is linked to the investigator-reported fraction f for a given sample j and cell type k by the following:

$$f'_{kj} = \alpha_k \beta_j f_{kj} \tag{3}$$

Here, α_k is a non-negative RNA source-specific scaling constant, and β_j is a non-negative sample-specific scaling constant. This formulation allows for straightforward correction of systematic reporting biases in f described earlier in the text. Of note, for this article, we do not constrain $\sum_k f'_{kj} = 1$ for all samples j, as normalization of microarray data during preprocessing typically results in small but computationally significant deviations from this ideal constraint. We do, however, include the option to constrain $\sum_k f'_{kj} = 1$ during implementation of MMAD.

2.3 Optimization

For microarray studies of mixed tissue specimens, the contributing cell types, their characteristic gene expression profiles, C, and the effective RNA fractions, f', may or may not be known. These parameters can be estimated in MMAD using a maximum likelihood approach by minimizing the residual sum of squares given by

$$\varepsilon = \sum_{i=1}^{I}\sum_{j=1}^{J} (x_{ij} - \hat{x}_{ij})^2 \tag{4}$$

where

$$\hat{X}_{ij} = \sum_{k=1}^{K} \hat{C}_{ik} \hat{f}'_{kj} \quad \text{and} \quad \hat{x}_{ij} = \log_2 \hat{X}_{ij} \tag{5}$$

Additional modifications are incorporated, depending on the level of prior knowledge about the cell types present in the mixed tissue.

2.3.1 C is known, f is unknown In this case, we optimize the unknown parameters by simply minimizing the residual sum of squares given by Equation (4) using a non-linear conjugate gradient search algorithm implemented in MATLAB. The maximum likelihood estimates of α, β and f are all constrained to be non-negative.

2.3.2 C is unknown, f' is known Estimation of C when f is known can be performed independently for each gene i by minimizing the residual sum of squares for that gene. We modify this approach by hypothesizing that some cell types may share common expression levels for gene i. To address this possibility and minimize overfitting, we evaluate all possible partitions of the K cell types into subsets, and constrain all cell types in a given subset to have identical expression levels for that gene i. Let P represent one possible partition, and let N_P be the number of distinct subsets under that partition, where $1 \le N_P \le K$. Let $\hat{C}^*_{is|P}$ be the expression level of gene i for cells in subset s under partition P. In this case, the estimated expression of gene i for cell k is given by

$$\hat{C}_{ik} = \sum_{s=1}^{N_P} \delta_{sk|P} \cdot \hat{C}^*_{is|P} \tag{6}$$

where

$$\delta_{sk|P} = \begin{cases} 1, & \text{if cell } k \in \text{subset } s, \text{ given partition } P \\ 0, & \text{else} \end{cases} \tag{7}$$

Let $\varepsilon_{i|P}$ be the residual sum of squares after optimizing model fit under partition P for gene i. To select which partition is best supported for each gene by the available data, we use the corrected Akaike information criterion (AIC_c) as defined by Hurvich and Tsai (1989). AIC_c provides a measure of model fit that balances the likelihood of observing the given data under the model against the number of parameters estimated for that model. The AIC_c for partition P and gene i is given by

$$\text{AIC}_\text{c} = J \cdot \log \varepsilon_{i|P} + \frac{2 \cdot J \cdot (N_P + 1)}{J - N_P - 2} \tag{8}$$

The partition that is associated with the lowest AIC_c for each gene is used to generate the final estimates of $\hat{C}$ for that gene. It is important to note that different partitions may be selected for different genes. In situations where it is computationally intractable to consider all possible partitions, we can approximate the number of distinct subsets by

$$N_P \approx \sum_{p=1}^{K}\left(\sum_{q=1}^{K} \exp\left[-\left(\frac{\hat{c}_{ip} - \hat{c}_{iq}}{\Delta_c}\right)^2\right]\right)^{-1} \tag{9}$$

where Δ_c is a small positive number. We solve for the values of $\hat{C}$ that minimize the approximate AIC_c. We gradually shrink $\Delta_c \to 0$, allowing the final estimates to approach an AIC_c-informed model fit. Of note, during optimization, we constrain our estimates $\hat{c}_{ik} = \log_2 \hat{C}_{ik}$ to fall approximately within the same dynamic range as the given dataset by using the minimum ($x_{\min}$) and maximum ($x_{\max}$) expression values over all genes and all samples in the dataset to define our constraint boundaries:

$$(x_{\min} - 1) \le \hat{c}_{ik} \le (x_{\max} + 1) \tag{10}$$

2.3.3 C and f are unknown In this case, we note that maximum likelihood estimates of C and f' obtained by minimizing [Equation (4)] are not uniquely determined without the incorporation of prior knowledge or additional constraints (Taslaman and Nilsson, 2012). We agree with Gaujoux and Seoighe (2012) and Zhong *et al.* (2013) that *a priori* knowledge about the cell types of interest should be incorporated if possible; in particular, if certain genes are known to be highly specific for a given cell type, then those genes can be used to estimate f for that cell type. Let us assume that for each cell k, there exists a subset of genes, G_k, which is highly specific to that cell k. Then, for all genes $g \in G_k$, x_{gj} would be approximated by

$$x_{gj} = \log_2(C_{gk} f_{kj} + B_g) + e_{ij} \tag{11}$$

where B_g is the background expression for gene g. Unfortunately, a specific list of genes is not always available for each of the cell types present in a given sample. For the purposes of MMAD, in the absence of prior information about the constituent cell types, we assume that those genes with the highest variability in the sample set are most likely to be differentially expressed by the different cell types. We assign each of these highly variable genes (default is the top 1% most variable genes) to a putative cell type using k-means clustering with Pearson's correlation coefficient as the distance metric.

Once appropriate cell-specific gene subsets have been established, we estimate C, f and B for each gene by minimizing the residual squared error.

$$\varepsilon_G = \sum_{k=1}^{K} \left[\sum_{g \in G_k} \sum_{j=1}^{J} (x_{gj} - \hat{x}_{gj})^2 \right] \tag{12}$$

We use the estimate of f obtained in this manner to approximate C for all genes in the sample by AIC_c optimization as outlined in Section 2.3.2.

2.4 Tests for differences between classes

Assume that the samples in a given dataset X can be divided into two distinct classes. Let $C_{(1)}$ and $C_{(2)}$ be the characteristic gene expression profiles for the different cell types, and let $f'_{(1)}$ and $f'_{(2)}$ be the class-specific cell fractions. Differences in gene expression and cell type composition between classes are assessed as follows:

2.4.1 Differences in cell fraction between classes Estimates of effective RNA fraction in each sample are first obtained using the appropriate algorithm mentioned earlier in the text without using class information. Class-specific cell fractions are then compared using a two-tailed unpaired t-test as implemented in MATLAB using the function *t-test2*().

2.4.2 Differences in cell type-specific gene expression between classes In cases where f is unknown, we first estimate f for all of the samples as per Section 2.3.3. For each gene i, we then compare the null model of differential expression (no differential expression in any of the cell types between the two classes) with all models in which gene i is differentially expressed by just one of the cell types k. For computational reasons, we do not consider models in which >1 cell type differentially expresses gene i. We calculate the AIC_c for each of these models, as well as the AIC_c weight, which is the relative support for the given model when compared with all models under consideration:

$$\omega_{\text{model[diff}(k)]} = \frac{e^{-\frac{1}{2}\text{AIC}_{\text{c, model[diff}(k)]}}}{e^{-\frac{1}{2}\text{AIC}_{\text{c, model[null]}}} + \sum_h e^{-\frac{1}{2}\text{AIC}_{\text{c, model[diff}(h)]}}} \tag{13}$$

We define the MMAD cell-specific differential expression test statistic for gene i and cell type by

$$\text{M}_{\text{cell-specific, } ik} = -\log\left(1 - \omega_{\text{model[diff}(k)]}\right) \tag{14}$$

Larger values are consistent with greater support for differential expression of gene i in cell type k. P-values and false discovery rates can be estimated using a permutation-based approach in which class labels are randomly permuted and the statistics are recalculated.

2.4.3 Global test for differences in cell-specific gene expression The global test for differences in cell-specific gene expression summarizes the existing support for differential expression in at least one of the cell types present in the sample:

$$\text{M}_{\text{global, } i} = -\log\left(1 - \sum\nolimits_k \omega_{\text{model[diff}(k)]}\right) \tag{15}$$

Once again, larger values are consistent with greater evidence for differential expression for gene i in at least one of the cell types in the sample. P-values and false discovery rates can be estimated using a permutation-based approach.

2.4.4 Comparison metrics for differences in gene expression As our comparison metrics, we evaluated (i) a two-tailed t-test, (ii) significance analysis of microarrays (SAM) (Tusher *et al.*, 2001) and (iii) cell type-specific significance analysis of microarrays (csSAM) (Shen-Orr *et al.*, 2010). The t-test was implemented using the MATLAB function *t-test2*() (MATLAB R2012b), SAM using the samr() package in R (version 2.15.1) and csSAM using the csSAM R package (R version 2.15.1).

2.5 Simulation

To assess model performance, we created a dataset consisting of simulated colorectal adenocarcinoma cells, adipose cells and $CD8^+$ T-cells in different mixing proportions. For each of the selected cell types, we used publicly available gene expression data from the Gene Expression Omnibus (GEO), series GSE1133, to define the characteristic gene expression profiles (Su *et al.*, 2004). All samples selected had been run on the Affymetrix HG U133A Array. We downloaded the raw CEL files from GEO and used the median-normalized gene expression data to define the cell-specific expression profiles (C). For each sample j and cell type k that was simulated, we sampled the effective RNA fractions (f'_{kj}) from a Dirichlet distribution. The final value of x_{ij} was calculated per Equation (1) with e_{ij} drawn from a normal distribution with variance set equal to the global variance in gene expression between duplicate samples in GSE1133.

2.6 Data preprocessing

Background adjustment for simulated data was performed using the MATLAB function *rmabackadj()*. Median and quantile normalization were performed using the MATLAB functions *manorm()* and *quantilenorm()*, respectively. Robust Multichip Average (RMA) summarization (Bolstad *et al.*, 2003) was performed using the MATLAB function *rmasummary()*. Prefiltering of Affymetrix probes was performed using the Micro Array Suite 5.0 (MAS5.0) algorithm with $\tau = 0$ (Hubbell *et al.*, 2002).

3 RESULTS

We present a method for the *in silico* microdissection of tissue samples into component cell types that is both flexible and robust.

3.1 Properties of MMAD deconvolution algorithm

3.1.1 Data preprocessing Background subtraction and quantile normalization can alter the linear relationship between probes in an intensity-dependent manner; median normalization preserves linearity though it is less robust when adjusting for

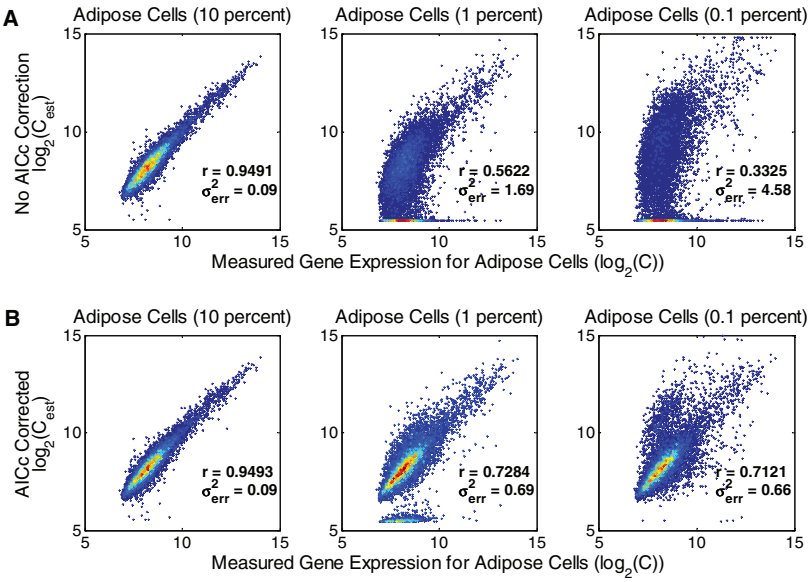

Fig. 1. AIC_c correction during model fit improves estimates of gene expression. We simulated three datasets containing mixtures of colon adenocarcinoma cells, $CD8^+$ T-cells and adipose cells in different mixing proportions. The average percentage of adipose cells was decreased in each simulation (10, 1, 0.1%, respectively). Without AIC_c correction, we note a marked decrease in the ability of MMAD to predict the gene expression profile of adipose cells when the fraction of adipose cells drops to 1% or below (**A**). AIC_c correction dramatically stabilizes predictions, even at cell fractions averaging 0.1 percent (**B**). (See also Supplementary Fig. S3.)

intensity-related bias. Importantly, choice of normalization method can directly influence the ability to detect differentially expressed genes (Chiogna *et al*., 2009). We compared MMAD performance with and without RMA background subtraction and compared median normalization with quantile normalization using simulated data consisting of 40 samples generated as per Section 2.5. We noted comparable performance using quantile normalization and median normalization, suggesting that these methods may be interchangeable for initial probe normalization; however, predicted differences in gene expression between cell types was poorer when background correction was performed before the model fit (Supplementary Fig. S2). Therefore, for high-quality datasets with stable signal-to-noise ratios, we recommend using median or quantile normalization and not background correcting before performing computational microdissection with MMAD. We use quantile normalization for the remainder of this article.

3.1.2 AIC_c adjustment improves the prediction of cell-specific gene expression Overfitting can be a major concern during deconvolution, particularly when the number of samples is small and when individual cell types are typically present at low frequencies in a given sample. We evaluated MMAD performance in three simulated datasets generated as per the methodology outlined in Section 2.5. Each dataset consisted of 10 samples; we decreased the relative percentage of adipose cells in each of the datasets, such that on average adipose cells comprised 10% of cells in the first dataset (70% colon, 20% T-cell), 1% of cells in the second dataset (79% colon, 20% T-cell) and 0.1% of cells in the third dataset (79.9% colon, 20% T-cell). We estimated gene expression profiles for each of the three cell types both with and without AIC_c correction using the known values of f. We note that there was a marked improvement in the model fit with AIC_c correction, particularly for adipose cells, which were present at the lowest frequency in the simulation (Fig. 1 and Supplementary Fig. S3).

3.2 Computational microdissection of tissue mixture datasets

Given model performance with the simulated datasets, we investigated performance on available tissue mixture datasets.

3.2.1 Estimation of cell-specific gene expression (C) We used a benchmark dataset consisting of pure brain, liver and lung tissue from a single rat in isolation as well as in 11 different mixture ratios as described in Shen-Orr *et al*. (2010). Each mixture ratio is characterized by 3 technical replicates with 42 total samples (9 pure tissue samples, 33 mixed tissue samples). RNA expression levels were measured with the Rat Genome 230 2.0 Array (Affymetrix). Raw data are available on the GEO site, GSE19830. CEL files were downloaded from GEO; probesets were prefiltered using the MAS5.0 detection algorithm with a threshold $P < 0.05$ in at least two samples, and data were quantile normalized and summarized using *rmasummary*().

Computational microdissection of the 33 mixed tissue specimens was performed using MMAD and compared with results from csSAM. MMAD was constrained to use the investigator supplied cell fractions f and allowed to fit the model using both a strict supervised fit (α and β constrained to equal 1) and a semi-supervised fit (α and β unconstrained). For csSAM, we used the function csfit() as implemented in the csSAM R package (R version 2.15.1). We performed the deconvolution in csSAM in both log-space (original implementation) and linear space based on

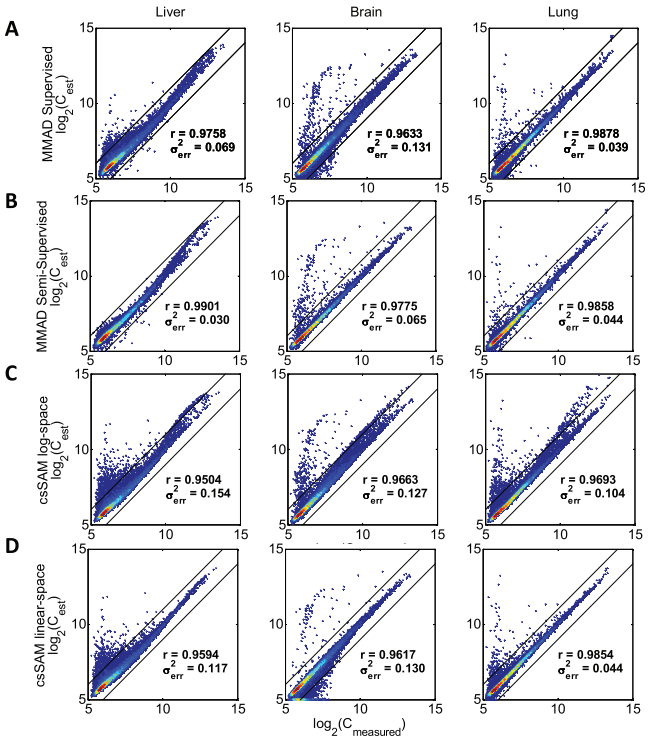

Fig. 2. Computational microdissection of rat tissue mixture dataset (GSE19830). We compared the gene expression of pure liver, brain and lung tissue with estimates obtained by deconvoluting impure (mixed) tissue samples using MMAD and csSAM. Both MMAD and csSAM were constrained to use the investigator supplied values of *f*. We evaluated performance of MMAD without normalization adjustments (supervised approach) (**A**) and with normalization adjustments (semi-supervised approach) (**B**). Results were compared with csSAM deconvolution in both log-space (**C**) and linear space (**D**). We note that MMAD outperforms csSAM in both log-space and linear space; in particular, the residual variance is markedly reduced when MMAD is run in a semi-supervised manner using the bias-correction parameters

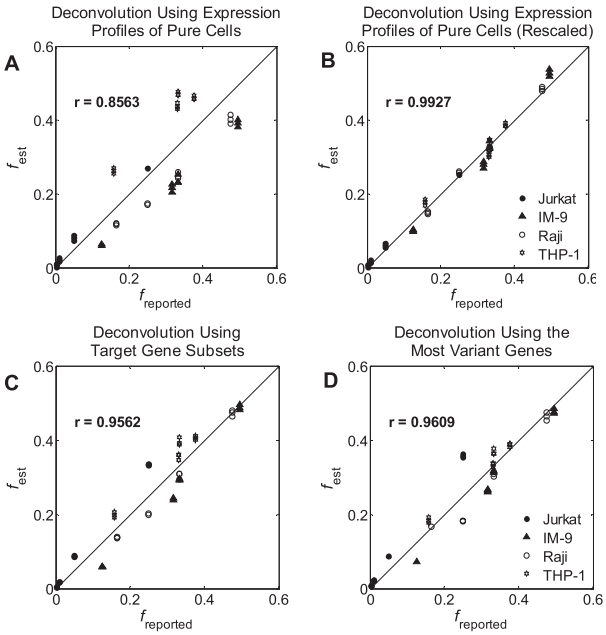

Fig. 3. Estimation of cell fraction using MMAD in an immune mixture dataset (GSE11058). We estimated the fractional contribution of individual immune cells to mixed samples using gene expression profiles of pure immune cells (**A**). There is a slight scaling artifact, which can be filtered out by multiplying the gene expression profiles of pure cells by an appropriate constant (normalization artifact) (**B**). Results are similarly robust using predefined characteristic gene subsets (**C**) or with a blind deconvolution using the top 1% most variable genes (**D**)

reports of potentially improved performance in linear space (Zhong and Liu, 2012).

The estimates of gene expression for each constituent tissue type were best when using MMAD with the semi-supervised fit ($r_{liver} = 0.99$, $r_{brain} = 0.98$, $r_{lung} = 0.99$). Though csSAM in linear space performed better than csSAM in log-space, there was still significant residual model error (σ^2_{err}), particularly for genes with low levels of expression in the different cell types (Fig. 2).

3.2.2 Estimation of cellular fraction (f,f′) We used a test dataset consisting of mixtures of four immune cell lines (Jurkat, IM-9, Raji and THP-1) that is available for public download from GEO (GSE11058) (Abbas *et al.*, 2009). Deconvolution of this particular dataset is challenging due to the fact that the representative cell lines have highly correlated gene expression signatures given their common immune phenotype. CEL files were downloaded from the GEO site, prefiltered with the MAS5.0 detection algorithm ($\tau = 0$) and kept for further analysis if a detection $P < 0.05$ was present in at least two samples. Data were then quantile normalized and summarized using *rmasummary()*. We evaluated three different approaches to the estimation of cell fraction in this dataset: (i) estimation of *f* (or *f′*) using the complete gene expression profiles of pure cells (*C*), (ii) estimation of *f* (or *f′*) using a subset of cell-specific marker genes and (iii) blind estimation of cell fraction without prior knowledge of constituent cell types. For our target gene subset in (ii), we used those genes that were expressed at least 5-fold higher in the target cell as compared with any of the other cell types. For the blind deconvolution in MMAD, we used the top 1% most variable genes in log-space as our target genes and assigned each of these genes to a putative cell type of interest using *k*-means clustering as implemented in MATLAB with the *kmeans*() algorithm. As comparators to MMAD, we also evaluated the performance of quadratic programing implemented in MATLAB as proposed by Gong *et al.* (2011), gene subset-guided deconvolution using the Digital Sorting Algorithm (DSA) algorithm proposed by Zhong *et al.* (2013) and blind deconvolution using the deconf() algorithm proposed by Repsilber *et al.* (2010) implemented in R (version 2.13.1).

For all MMAD-derived estimates, there was a tight correlation between estimates of *f* (or *f′*) and reported cell fractions (Figs 3 and 4). The blind deconvolution approach was particularly impressive for MMAD ($r = 0.961$) and outperformed the comparator ($r = 0.798$). We did note, however, that when the gene expression profiles of pure cells were used to computationally microdissect the mixed samples (approach 1) that there was a scaling artifact, which could be resolved by rescaling each cell type by a constant term. This artifact was also noted for the comparator method, supporting the notion of either normalization bias or reporting bias.

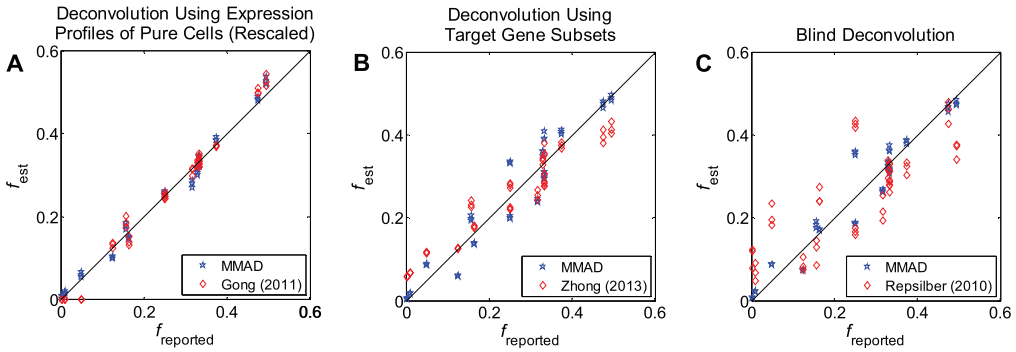

Fig. 4. Comparison of approaches for estimation of cell fraction. MMAD compares favorably with all comparators. We note comparable performance between MMAD and quadratic deconvolution (Gong *et al.*, 2011) after rescaling (renormalizing) the reported gene expression of pure cell types ($r > 0.99$ for both approaches). Performance is also comparable for deconvolution with target gene subsets between MMAD and DSA (Zhong *et al.*, 2013). MMAD outperforms deconf() (Repsilber *et al.*, 2010) in blind deconvolution

3.3 Deconvolution with MMAD improves detection of differentially expressed genes

We then proceeded to investigate whether our model was able to improve the detection of differential expression of genes that are differentially expressed only in a specific cell type.

3.3.1 Simulation We considered a mixed tissue dataset in which colon adenocarcinoma cells, CD8+ T-cells, and adipose cells are present in varying mixing proportions as per Section 2.5. We divided these samples into two distinct classes. For samples in class 1, we downregulated 1% of all genes in each of the cell types by a factor of 1.5. We also up regulated 1% of all genes in each cell type by a factor of 1.5. These genes were selected at random. A total of 40 samples were simulated for each dataset with 20 samples in class 1 and 20 samples in class 2.

3.3.2 Performance of the global MMAD test statistic We hypothesized that improvements in the detection of differential gene expression using the global MMAD test statistic would be most noticeable in samples with highly variable cell type composition. To assess this, we simulated three datasets as per Section 3.3.1 with high, moderate and low variability in cell fraction (Dirichlet concentration parameters: 3, 30 and 300, Supplementary Fig. S4). In all cases, the expected fractions of colorectal adenocarcinoma cells, $CD8^+$ T-cells and adipose cells were fixed at 0.7, 0.2 and 0.1. We compared the global MMAD test statistic versus the SAM test statistic for the detection of differentially expressed genes. Receiver-Operating Characteristic curves are summarized in Figure 5A. We note that the performance of MMAD is superior to SAM in cases where cell type fraction is at least moderately variable. In cases where cell type variability is low, performance is comparable between MMAD and SAM.

3.3.3 Performance of the cell-specific MMAD test statistic We then evaluated the ability of MMAD to properly attribute differences in gene expression to the appropriate cell type (Dirichlet concentration parameter 25, expected cell fractions unchanged from Section 3.3.2). We computed the cell-specific MMAD test statistic for each gene and each cell type as well as the csSAM cell-specific differential expression test statistic. MMAD performance was similar to csSAM for the most highly expressed cell type (colon adenocarcinoma). However, for both the $CD8^+$ T-cells and the adipose cells, detection of differential expression was superior with MMAD (Fig. 5B).

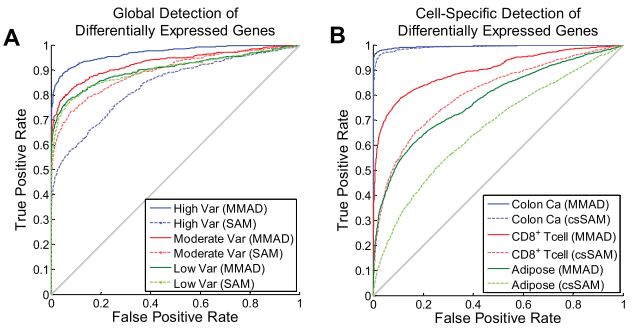

Fig. 5. Detection of differentially expressed genes is improved with MMAD. The global MMAD test statistic provides greater discriminatory power than the SAM test-statistic for the detection of differentially expressed genes in mixed tissue samples with highly variable and moderately variable cell type composition; differences are not seen when variability in cell type fraction is low (**A**). In a moderately variable simulation, MMAD and csSAM perform similarly for detecting differences in differential expression for the major cell type present in the simulated mixed tissue samples (colon adenocarcinoma cells, average cell fraction 0.7); MMAD shows improved ability to detect differential expression in cell types present at lower frequencies in the simulation ($CD8^+$ T-cells and adipose cells, average cell fractions 0.2 and 0.1) (**B**)

4 DISCUSSION

Numerous studies have demonstrated the potential of microarray expression profiling as a diagnostic, prognostic and predictive tool. However, widespread adoption of microarray technologies into clinical and laboratory practice has been limited in part by the problem of tissue heterogeneity, which is an unavoidable challenge for investigators evaluating real-world samples. Such heterogeneity can mask key differences between sample classes and can limit the ability of investigators to generalize results across studies and across different institutions.

In this article, we designed and validated a flexible new method for performing computational microdissection of complex tissues and analyzing differences between classes. By accounting for this heterogeneity with robust methodologies, we believe that investigators will improve their ability to discover pertinent (and potentially disease-modifying) differences in gene expression.

MMAD incorporates features of previous algorithmic approaches to tissue deconvolution (Abbas *et al*., 2009; Bolen *et al*., 2011; Erkkila *et al*., 2010; Gong *et al*., 2011; Lahdesmaki *et al*., 2005; Lu *et al*., 2003; Repsilber *et al*., 2010; Shen-Orr *et al*., 2010; Stuart *et al*., 2004; Venet *et al*., 2001; Wang *et al*., 2006) and also introduces several novel ideas. Key features of MMAD include the following:

(1) Computational microdissection with MMAD does not require prior knowledge about contributing cell types (although such knowledge can be incorporated into the model). We have demonstrated that estimates of cell fraction and cell type-specific gene expression can be obtained using existing gene expression data or cell fraction data when available or by using a 'blind' deconvolution. In all cases, results obtained with MMAD compare well to previously described approaches. This is particularly true for 'blind' deconvolutions. The flexibility of this approach is important for analysis of many biologic systems, as the fractional contribution of individual cell types may be unknown or known with limited precision.

(2) Biologic variation is modeled assuming log-normal variation, an assumption which is shared with several standard normalizations and analysis metrics used to test for differential expression (Bolstad *et al*., 2003; Tusher *et al*., 2001). Importantly, though, the deconvolution uses the natural scale gene expression values (not log-transformed) as recommended in Zhong and Liu (2012). This allows us to fit the observed data with more realistic models and, consequently, to make more natural inferences about differential gene expression. A similar statistical model was recently used by Ahn *et al*. (2013).

(3) We introduce the concept of effective RNA fraction (as distinct from the reported RNA fraction) that improves the results of computational microdissection by accounting for subtle normalization and reporting biases.

(4) We note that overfitting is a potential source of bias in computational microdissection, which has been largely overlooked in previous approaches. We include a standard information metric (AIC_c) during model fit to reduce potential overfitting. We demonstrate that the effect of this approach is most noticeable in cell types present at low frequencies within mixed tissue samples.

(5) We introduce several techniques that can be used to compare differences in gene expression between classes, including global differences in gene expression between sample classes (adjusted for tissue heterogeneity), differences in cell type frequency between sample classes and differences in gene expression between sample classes at the level of individual cell types. These metrics improve on standard approaches for the detection of differentially expressed genes (e.g. SAM) as well as more recently described approaches that attempt to incorporate computational microdissection (csSAM).

We note that deconvolution of gene expression data from RNAseq experiments is now an area of active research (Gong and Szustakowski, 2013). We note that although MMAD cannot be used directly for the analysis of RNAseq data, the approach that we have outlined in MMAD can easily be adapted for use in RNAseq experiments and other quantitative sequencing datasets by incorporating the appropriate statistical models. This would be a natural extension of our current work.

We believe the flexibility provided by MMAD will be key to allowing tissue deconvolution to be adopted more generally in microarray studies. Of particular interest will be the application of deconvolution to tumor samples with the potential to investigate biologically significant changes in both tumor and stroma. This concept has been strongly driven by findings that gene expression in tumor associated fibroblasts affects tumor growth in epithelial cells (Trimboli *et al*., 2009). Because MMAD has the potential to derive the characteristic gene expression of a specific cell type in the microenvironment of a tumor, or other tissue, we propose that it will now be possible to more rigorously evaluate specific contributions of the microenvironment to disease development in mixed tissue samples, including the evaluation of changes in gene expression in stromal cells, such as fibroblasts or macrophages.

Funding: Ruth L. Kirschstein T32 Institutional National Research Service Award (NRSA) (to D.A.L.) through the Department of Health and Human Services (5T32CA009338-33 and 5T32CA009338-34) and by NIH grant (5R01CA141090 to K.H. and J.D.P.).

Conflict of Interest: none declared.

REFERENCES

Abbas,A.R. *et al*. (2009) Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One*, **4**, e6098.

Ahn,J. *et al*. (2013) Demix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, **29**, 1865–1871.

Alizadeh,A.A. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Beer,D.G. *et al*. (2002) gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med*., **8**, 816–824.

Belbin,T.J. *et al*. (2002) Molecular classification of head and neck squamous cell carcinoma using cDNA microarrays. *Cancer Res*., **62**, 1184–1190.

Bittner,M. *et al*. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Bolen,C.R. *et al*. (2011) Cell subset prediction for blood genomic studies. *BMC Bioinformatics*, **12**, 258.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Chiogna,M. *et al*. (2009) A comparison on effects of normalisations in the detection of differentially expressed genes. *BMC Bioinformatics*, **10**, 61.

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.

Debey,S. *et al*. (2004) Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on

overall expression and the role of different cell types. *Pharmacogenomics J.*, **4**, 193–207.
Dhanasekaran,S.M. *et al.* (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature*, **412**, 822–826.
Elloumi,F. *et al.* (2011) Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. *BMC Med. Genomics*, **4**, 54.
Erkkila,T. *et al.* (2010) Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics*, **26**, 2571–2577.
Feezor,R.J. *et al.* (2004) Whole blood and leukocyte rna isolation for gene expression analyses. *Physiol. Genomics*, **19**, 247–254.
Gaujoux,R. and Seoighe,C. (2012) Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. *Infect. Genet. Evol.*, **12**, 913–921.
Gaujoux,R. and Seoighe,C. (2013) Cellmix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, **29**, 2211–2212.
Gong,T. *et al.* (2011) Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS One*, **6**, e27156.
Gong,T. and Szustakowski,J.D. (2013) Deconrnaseq: a statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data. *Bioinformatics*, **29**, 1083–1085.
Hubbell,E. *et al.* (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585–1592.
Hurvich,C.M. and Tsai,C.L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.
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.
Meyniel,J.P. *et al.* (2010) A genomic and transcriptomic approach for a differential diagnosis between primary and secondary ovarian carcinomas in patients with a previous history of breast cancer. *BMC Cancer*, **10**, 222.
Perou,C.M. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.
Repsilber,D. *et al.* (2010) Biomarker discovery in heterogeneous tissue samples - taking the in-silico deconfounding approach. *BMC Bioinformatics*, **11**, 27.
Shen-Orr,S.S. *et al.* (2010) Cell type-specific gene expression differences in complex tissues. *Nat. Methods*, **7**, 287–289.
Sorlie,T. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.
Stratford,J.K. *et al.* (2010) A six-gene signature predicts survival of patients with localized pancreatic ductal adenocarcinoma. *PLoS Med.*, **7**, e1000307.
Stuart,R.O. *et al.* (2004) *In silico* dissection of cell-type-associated patterns of gene expression in prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 615–620.
Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
Taslaman,L. and Nilsson,B. (2012) A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data. *PLoS One*, **7**, e46331.
Trimboli,A.J. *et al.* (2009) Pten in stromal fibroblasts suppresses mammary epithelial tumours. *Nature*, **461**, 1084–1091.
Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
van 't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
Venet,D. *et al.* (2001) Separation of samples into their constituents using gene expression data. *Bioinformatics*, **17 (suppl. 1)**, S279–S287.
Wang,M. *et al.* (2006) Computational expression deconvolution in a complex mammalian organ. *BMC Bioinformatics*, **7**, 328.
Zhong,Y. and Liu,Z. (2012) Gene expression deconvolution in linear space. *Nat. Methods*, **9**, 8–9.
Zhong,Y. *et al.* (2013) Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, **14**, 89.