Genome Medicine

COMMENTARY

Gene expression deconvolution in clinical samples

Yingdong Zhao and Richard Simon*

Abstract

Cell type heterogeneity may have a substantial effect on gene expression profiling of human tissue. Several *in silico* methods for deconvoluting a gene expression profile into cell-type-specific subprofiles have been published but not widely used. Here, we consider recent methods and the experimental validations available for them. Shen-Orr *et al.* recently developed an approach called cell-type-specific significance analysis of microarray for deconvoluting gene expression. This method requires the measurement of the proportion of each cell type in each sample and the expression profiles of the heterogeneous samples. It determines how gene expression varies among pre-defined phenotypes for each cell type. Gene expression can vary substantially among cell types and sample heterogeneity can mask the identification of biologically important phenotypic correlations. Consequently, the deconvolution approach can be useful in the analysis of mixtures of cell populations in clinical samples.

Background

Microarray expression profiling has proven to be a valuable technology in a wide variety of biological and biomedical investigations. One of its limitations, however, is the relatively large amount of mRNA required. Consequently, for analyses involving tissue from humans or experimental animals, the tissue samples used for mRNA extraction are often heterogeneous with regard to cell type. Because gene expression can vary substantially among cell types, gene expression profiles based on tissue samples of varying composition can be very difficult to interpret biologically. The problem is particularly serious for expression profiles intended for clinical use in informing treatment selection. Investigators have reported difficulties caused by sample heterogeneity for identifying biologically relevant differentially expressed genes and for developing and validating predictive models [1-3]. Although laser capture microdissection provides an experimental means for selecting a more homogeneous population of cells, it is time consuming and difficult to obtain sufficient purified tissue with adequately preserved RNA.

Expression deconvolution

Several statistical approaches have been proposed to deconvolute gene expression profiles obtained from heterogeneous tissue samples into cell-type-specific subprofiles. Most of the methods are based on a framework first proposed by Venet *et al.* [4], incorporating the linearity assumption that the expression of each gene in a mixture of cell types is a weighted average of the expression values that would exist for pure populations of those cell types. The weights are determined by the proportional composition of the cell types in the mixture and hence are the same for each gene but differ among sample mixtures. Since the publication of Venet *et al.* [4], several additional publications have appeared dealing with deconvolution of gene expression profiles on complex tissues (for example, [5-10]). Without reviewing the details that distinguish the various methods, we attempt here to summarize the status of this area of development.

When the proportions of the cell types in each mixture sample are known from fluorescence activated cell sorting analysis, histopathological evaluation or other experimental methods, deconvolution is relatively straightforward. With the known proportions of the cell types in the mixture, deconvolution can be solved as a linear regression problem in which the cell-type-specific gene expression levels represent the regression coefficients. In fact, under these conditions, the regression problem can be solved separately for each gene.

In some cases the cell-type-specific gene expression levels may be of interest in their own right, or interest may focus on differences in expression among cell types. For cancer studies, however, interest is often on differential expression among classes of tumors (such as

*Correspondence: rsimon@mail.nih.gov
Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA

responders versus non-responders to a treatment), with expression from normal epithelium and infiltrating immune cells of lesser interest. Shen-Orr *et al.* [8] developed cell-type-specific significance analysis of microarray (csSAM) for analyzing differentially expressed genes for each cell type in sample mixtures with microarray data. The relationship between measured gene expression in mixed samples and the expression of genes in the isolated pure subsets was tested experimentally for synthetic mixtures of liver, brain and lung cells from rats. Their *in silico* synthesized mixture expression profiles, obtained by multiplying the measured pure tissue expression profiles by the proportion of the tissue subset in a given mixture sample, were highly correlated with the experimentally measured expression profiles for the mixtures. This provided direct support for the linearity assumption of all previous models. The deconvoluted estimates of cell-type-specific expression were in good agreement with expression measured in pure cell types for the vast majority of probes.

The authors [8] then applied csSAM to human whole blood gene expression array data from kidney transplant recipients. When they used the whole blood analyses, there were no differentially expressed genes detected between the rejection group and stable group. However, a large number of differentially expressed genes were identified between the two groups in two individual cell types when applying the csSAM for each of the five quantified cell types: monocyte, basophile, neutrophil, eosinophil and lymphocyte. The method requires experimental measurements of the proportional composition of the component cell types in each sample. Although there are some pre-processing issues such as normalization that require further consideration, csSAM seems to be a useful tool for analysis of gene expression profiling of heterogeneous samples with known relative cell type frequencies. Source code for csSAM in the R statistical programming language is available [8].

Several investigations performed deconvolution when the proportions of the component cell types were unknown but expression of signature genes in pure cell types was known (for example, [5-7]). Abbas *et al.* [7] developed an approach to estimate the proportions of white blood cell subtypes in samples from patients with systemic lupus erythematosus. First, they selected the most highly expressed signature probesets (genes) among several of the 18 immune cell types of interest using the expression data from the pure cells. They then used expression profiles for these signature genes to solve a linear equation for the proportions of the 18 immune cell subtypes in both healthy donors and patients with lupus. The deconvoluted results allowed them to find patterns of leukocyte dynamics and their correlations with clinical outcomes. In circumstances such as described by Abbas *et al.* [7] in which careful preliminary studies have been conducted to identify signature genes and determine their expression in pure cell subtypes, such deconvolution can be successful.

Some proposals for deconvolution have been made for cases in which neither the proportions of the cell types in the mixtures nor signature genes are known. These approaches use a variety of methods, such as non-negative matrix factorization [9,10]. The validations available are limited, however, and the number of samples required for accurate deconvolution may be large [9]. Consequently, when measurements of the proportions of the component cell types in individual samples are not available and signature genes for each cell subtype are unknown, we believe that the status of deconvolution of expression profiles of mixtures is less clear.

Identifying genes that are differentially expressed among groups of diseased tissue samples is a frequent objective of gene expression profiling. Many of the publications referenced here ignore class information (such as disease versus normal or responder versus non-responder) in performing the deconvolution and state or imply that the deconvoluted cell-type-specific expression profiles can then be used with standard software packages for investigating class comparisons [6,10]. This approach is potentially problematic, however, because the deconvoluted expression profiles are no longer statistically independent. Shen-Orr *et al.* [8] indicate that the deconvolution should be performed separately for each class being compared and that in using permutation tests to assess statistical significance, deconvolution should be repeated for each permutation of class labels.

Conclusions

Deconvolution of gene expression profiles for heterogeneous samples can be performed accurately when sufficiently accurate estimates of the proportional representation of component cell types in each sample are available and when expression profiles of the components are sufficiently different. The csSAM method developed by Shen-Orr *et al.* [8] can be useful in such clinical applications. Further studies are needed to address potential confounding factors for deconvolution, such as data normalization and batch effect adjustment. As Shen-Orr *et al.* [8] indicated, although the assumption of linearity holds for majority of probes, identification and exclusion of probes affected by non-linear amplification or synergistic cross-hybridization may provide more accurate deconvolution. Although most of the previous deconvolution methods have focused on single-label microarray data, they could be potentially adapted for use with dual-label array that uses a homogeneous reference sample.

Deconvolution of expression profiles when estimates of the proportional representation of component cell types in each sample are not available can be performed accurately in cases, such as that of Abbas *et al.* [7], in which careful preliminary studies have been conducted to identify expression profiles of signature genes from pure samples that clearly distinguish the cell types. Without the prior identification of such signature genes or the measurement of cell-type proportions, however, methods for the deconvolution of gene expression profiles for mixed tissue samples require further investigation and experimental validation to clarify the conditions under which accurate results can be obtained.

Abbreviations
csSAM, cell-type-specific significance analysis of microarray.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Both authors contributed equally to this article.

Published: 29 December 2010

References

1. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO: **Individuality and variation in gene expression patterns in human blood.** *Proc Natl Acad Sci USA* 2003, **100:**1896-1901.
2. Michiels S, Koscielny S, Hill C: **Random effects survival models gave a better understanding of heterogeneity in individual patient data meta-analyses.** *J Clin Epidemiol* 2005, **58:**238-248.
3. Cleator SJ, Powles TJ, Dexter T, Fulford L, Mackay A, Smith IE, Valgeirsson H, Ashworth A, Dowsett M: **The effect of the stromal component of breast tumours on prediction of clinical outcome using gene expression microarray analysis.** *Breast Cancer Res* 2006, **8:**R32.
4. Venet D, Pecasse F, Maenhaut C, Bersini H: **Separation of samples into their constituents using gene expression data.** *Bioinformatics* 2001, **17 Suppl 1:**S279-S287.
5. Lu P, Nakorchevskiy A, Marcotte EM: **Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations.** *Proc Natl Acad Sci USA* 2003, **100:**10370-10375.
6. Wang M, Master SR, Chodosh LA: **Computational expression deconvolution in a complex mammalian organ.** *BMC Bioinformatics* 2006, **7:**328.
7. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF: **Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus.** *PLoS ONE* 2009, **4:**e6098.
8. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ: **Cell type-specific gene expression differences in complex tissues.** *Nat Methods* 2010, **7:**287-289.
9. Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J, Parida SK, Kaufmann SH, Jacobsen M: **Biomarker discovery in heterogeneous tissue samples - taking the in-silico deconfounding approach.** *BMC Bioinformatics* 2010, **11:**27.
10. Lähdesmäki H, Shmulevich L, Dunmire V, Yli-Harja O, Zhang W: **In silico microdissection of microarray data from heterogeneous cell populations.** *BMC Bioinformatics* 2005, **6:**54.

doi:10.1186/gm214
Cite this article as: Zhao Y, Simon R: **Gene expression deconvolution in clinical samples.** *Genome Medicine* 2010, **2:**93.