BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 1 2015, pages 137–139
doi:10.1093/bioinformatics/btu607

Data and text mining Advance Access publication September 10, 2014

UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples

Niya Wang[1,†], Ting Gong[2,†], Robert Clarke[3], Lulu Chen[1], Ie-Ming Shih[4], Zhen Zhang[4], Douglas A. Levine[5], Jianhua Xuan[1] and Yue Wang[1,*]

[1]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, [2]Department of Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center, Smithville, TX 78957, [3]Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20057, [4]Departments of Pathology and Oncology, Johns Hopkins University, Baltimore, MD 21231 and [5]Department of Surgery, Memorial Sloan-Kettering Cancer Center, New York, NY 10021, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: We develop a novel unsupervised deconvolution method, within a well-grounded mathematical framework, to dissect mixed gene expressions in heterogeneous tumor samples. We implement an R package, UNsupervised DecOnvolution (UNDO), that can be used to automatically detect cell-specific marker genes (MGs) located on the scatter radii of mixed gene expressions, estimate cellular proportions in each sample and deconvolute mixed expressions into cell-specific expression profiles. We demonstrate the performance of UNDO over a wide range of tumor–stroma mixing proportions, validate UNDO on various biologically mixed benchmark gene expression datasets and further estimate tumor purity in TCGA/CPTAC datasets. The highly accurate deconvolution results obtained suggest not only the existence of cell-specific MGs but also UNDO's ability to detect them blindly and correctly. Although the principal application here involves microarray gene expressions, our methodology can be readily applied to other types of quantitative molecular profiling data.

Availability and implementation: UNDO is available at http://bioconductor.org/packages.

Contact: yuewang@vt.edu

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

Received on May 6, 2014; revised on August 8, 2014; accepted on September 7, 2014

1 INTRODUCTION

Tumor–stroma interactions serve as both a major confounding factor and an underexploited information source in studying tumor and microenvironment (Junttila and de Sauvage, 2013). Although analyzing tumor cells in their microenvironment provides the most relevant context, mixed expressions cannot be resolved directly by global profiling (Clarke *et al.*, 2008). Experimental solutions to isolate pure cells are inconvenient and have many limitations (Hoffman *et al.*, 2004). Current computational alternatives perform basically a supervised deconvolution (Ahn *et al.*, 2013; Clarke *et al.*, 2010; Gosink *et al.*, 2007; Yoshihara *et al.*, 2013), where the required a priori information is often inaccurate or context dependent, thus greatly limits supervised approaches.

Supported by a well-grounded mathematical framework, we argue that both cell-specific expression profiles and mixing proportions can be estimated in a completely unsupervised mode from two or more heterogeneous samples using raw measured gene expression values. Fundamental to the success of our approach is the geometric identifiability of cell-specific marker genes (MGs) warranted by (i) non-negativity of gene expression values and (ii) co-definition of distinct phenotypes and cell-specific MGs.

The UNsupervised DecOnvolution (UNDO) R package adapts and extends recent unsupervised deconvolution framework in the literature (Chen *et al.*, 2011). Using UNDO to dissect tumor–stroma mixed gene expressions, we show that (Fig. 1) (i) the scatterplot of mixed cell expression profiles is a compressed version of the scatterplot of pure cell expression profiles; (ii) resident genes on the two radii of scatter sector are cell-specific MGs, and furthermore, the two radius vectors defined by the MGs coincide with mixing proportions. Accordingly, UNDO first detects MGs on the two radii of the scatter sector in tumor samples, then estimates cell proportions using standardized average expression values of MGs and finally uncovers pure cell expression profiles by matrix inversion. We demonstrate the performance of UNDO on both synthetic and benchmark real gene expression datasets with highly accurate deconvolution results. We further apply UNDO to estimate tumor purity in three datasets from TCGA (The Cancer Genome Atlas)/CPTAC (Clinical Proteomic Tumor Analysis Consortium) and obtain highly comparable results with the estimates by BACOM 2.0 and ABSOLUTE based on somatic copy number data (Carter *et al.*, 2012; Yu *et al.*, 2011).

2 DESCRIPTION

2.1 Methods and software

We adopt the linear latent variable model of raw measured expression data, given by (bold font indicates column vectors),

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

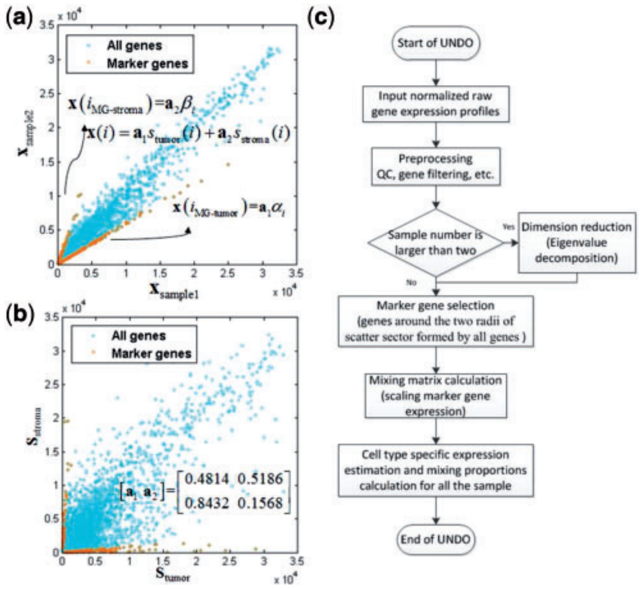

Fig. 1. (**a**) The scatterplot of mixing gene expression profiles (MCF7 and HS27 cell lines), which forms a scatter sector (a sector-shaped distribution). (**b**) The scatterplot of recovered pure cell gene expression profiles (MCF7 and HS27 cell lines). (**c**) The flowchart of UNDO algorithm

for genes $i = 1, \ldots, n$,

$$\begin{bmatrix} x_{\text{sample1}}(i) \\ x_{\text{sample2}}(i) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} s_{\text{tumor}}(i) \\ s_{\text{stroma}}(i) \end{bmatrix} \rightarrow \mathbf{x}(i) \qquad (1)$$
$$= \mathbf{a}_1 s_{\text{tumor}}(i) + \mathbf{a}_2 s_{\text{stroma}}(i),$$

where $s_{\text{tumor}}(i)$ and $s_{\text{stroma}}(i)$ are the gene expression values in pure cells, $x_{\text{sample1}}(i)$ and $x_{\text{sample2}}(i)$ are the gene expression values in heterogeneous samples and a_{jk} are the mixing proportions with $a_{11} + a_{12} = a_{21} + a_{22}$ (after signal normalization). Our method is built on a linear mixture of normalized raw measured data without log transformation (Ahn *et al.*, 2013). We further adopt the concept of cell-specific MGs (Yoshihara *et al.*, 2013), i.e. genes whose expressions are exclusively enriched in a particular cell population, or mathematically $\mathbf{s}(i_{\text{MG-tumor}}) \approx [\alpha_i\ 0]^T$ and $\mathbf{s}(i_{\text{MG-stroma}}) \approx [0\ \beta_i]^T$. As raw measured gene expression values are non-negative, when cell-specific MGs exist for each cell type, the linear latent variable model (1) is identifiable using two or more mixed expression profiles, supported by the following newly proved theorems:

Theorem 1 (scatter compression). *Suppose that pure cell expressions are non-negative and* $\mathbf{x}(i) = \mathbf{a}_1 s_{\text{tumor}}(i) + \mathbf{a}_2 s_{\text{stroma}}(i)$ *where* $\mathbf{a}_1$ *and* $\mathbf{a}_2$ *are linearly independent, then, the scatterplot of mixed expressions is compressed into a scatter sector (Fig. 1a, the set of light blue cross) whose two radii coincide with* $\mathbf{a}_1$ *and* $\mathbf{a}_2$.

Theorem 2 (unsupervised identifiability). *Suppose that pure cell expressions are non-negative and cell-specific MGs exist for each constituting cell type, and* $\mathbf{x}(i) = \mathbf{a}_1 s_{\text{tumor}}(i) + \mathbf{a}_2 s_{\text{stroma}}(i)$ *where* $\mathbf{a}_1$ *and* $\mathbf{a}_2$ *are linearly independent, then, the two scatter sector radii* $\mathbf{a}_1$ *and* $\mathbf{a}_2$ *of mixed expressions can be readily estimated from marker gene expression values.*

(See Supplementary Information for the formal proofs.)

Supported by the newly proved theorems, the UNDO algorithm performs the following major steps (Fig. 1c):

(1) preprocessing: quality control and removal of the minimally expressed genes whose norm is less than a pre-fixed positive small real number and outlier genes whose norm is bigger than a pre-fixed positive large real number on normalized raw data;

(2) dimension reduction when the sample number is larger than two;

(3) marker gene detection: identify the indices of cell-specific MGs located around the two radii of scatter sector that correspond to the genes with minimum/maximum ratio between the two mixed samples (Fig. 1a–1b, the set of orange diamond);

(4) estimate tumor–stroma proportions using marker gene expressions;

$$\hat{\mathbf{a}}_1 = \frac{1}{n_{\text{MG}-\text{tumor}}} \sum_{i \in \text{MG}-\text{tumor}} \frac{\mathbf{x}(i)}{\|\mathbf{x}(i)\|},$$
$$\hat{\mathbf{a}}_2 = \frac{1}{n_{\text{MG}-\text{stroma}}} \sum_{i \in \text{MG}-\text{stroma}} \frac{\mathbf{x}(i)}{\|\mathbf{x}(i)\|} \qquad (2)$$

where MG-tumor and MG-stroma are the index sets of MGs, and $n_{\text{MG-tumor}}$ and $n_{\text{MG-stroma}}$ are the numbers of MGs, for tumor and stroma, respectively; and ||.|| denotes the vector norm;

(5) estimate cell-specific expression profiles using matrix inversion.

More details on UNDO method, algorithm, parameter settings and alternative schemes are given in Supplementary Information.

2.2 Experimental validation and case study

We use five complementary evaluation criteria and the ground truth to assess the performance of UNDO method and algorithm. To assess the accuracy of cell proportion estimates, we use both Pearson correlation coefficients and $E1$ index (Moreau, 2001). To evaluate the accuracy of the estimated cell-specific expression profiles, we calculate the Pearson and concordance correlation coefficient (r_p *and* r_c) between the estimated expression profile and ground truth over both 'marker genes' (correct and rigorous way) and 'all genes' (can be misleading due to significant number of housekeeping genes). To assess the match/mismatch of membership/rank between the MGs detected from pure versus mixed expressions, we use Venn diagrams and Spearman's rank correlation coefficient.

To validate UNDO method and algorithm, we reconstituted tumor–stroma mixed expressions by multiplying pure expressions by pre-designed proportions. Solely using mixed expressions, UNDO accurately estimated the MGs, mixing proportions with $r_p = 0.99$, cell-specific expression profiles with $r_p = 0.99$ and $r_c = 0.99$ (see Supplementary Information).

We then tested UNDO algorithm on biologically mixed expressions from two breast cancer cell lines (MCF7-tumor

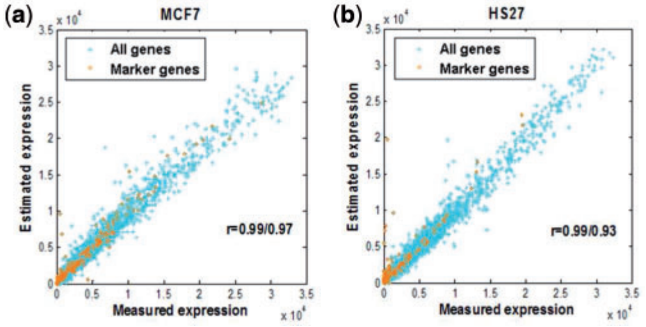

Fig. 2. (**a**) The scatterplot of the estimated versus true gene expression profile of MCF7 cell line. (**b**) The scatterplot of the estimated versus true gene expression profile of HS27 cell line

and HS27-stroma). The mRNA extracted from the individual cell lines are mixed with pre-specified proportions. Such mixtures mimic the actual biological samples with varying abundances of the constituent subsets from one another. UNDO method again accurately estimated the cell mixing proportions with $E1 = 0.0778$ ($r_p = 0.99$) and cell-specific expression profiles with average $r_p = 0.99$ and $r_c = 0.98$ between the deconvoluted expression profile and the measured expression profile in the pure cell lines (Fig. 2). The highly accurate deconvolution results obtained suggest not only the existence of cell-specific MGs but also UNDO's ability to detect them blindly and correctly.

We further assess UNDO's ability to detect differentially expressed genes (DEGs) without deconvolution. We compared the ranked DEGs indices detected by UNDO directly from mixed expressions, with the 'gold standard' reference DEGs identified from pure cell expressions, using Venn diagram, Spearman's rank correlation coefficient ($r_{\text{rank}} = 0.92$) and *receiver-operating characteristic* curve (AUC = 0.85). See Supplementary Information for more results on testing UNDO against ground truth and the estimates by BACOM 2.0 and ABSOLUTE on benchmark and TCGA/CPTAC datasets.

3 DISCUSSION

The UNDO software delivers a completely unsupervised deconvolution method for dissecting tumor–stroma mixed gene expressions (Supplementary Table S4). Tested on many benchmark datasets, UNDO is effective at detecting cell-specific MGs and DEGs and estimating cell proportions and cell-specific expression profiles. We expect UNDO method, with a Bioconductor R package, to be a useful tool for extracting cell-specific molecular signals in studying tumor–stroma interactions in many biological contexts. Though the UNDO method currently works for two-source mixtures only, it is principally applicable to the situation where tumor or non-tumor tissue is assumed to have a common composition of cellular subtypes across the tumor samples.

Funding: This work was supported by National Institutes of Health [Grant numbers NS29525, CA149147, CA160036, HL111362], in part.

Conflict of interest: none declared.

REFERENCES

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

Carter,S.L. *et al*. (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol*., **30**, 413–421.

Chen,L. *et al*. (2011) Tissue-specific compartmental analysis for dynamic contrast-enhanced MR imaging of complex tumors. *IEEE Trans. Med. Imaging*, **30**, 2044–2058.

Clarke,J. *et al*. (2010) Statistical expression deconvolution from mixed tissue samples. *Bioinformatics*, **26**, 1043–1049.

Clarke,R. *et al*. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat. Rev. Cancer*, **8**, 37–49.

Gosink,M.M. *et al*. (2007) Electronically subtracting expression patterns from a mixed cell population. *Bioinformatics*, **23**, 3328–3334.

Hoffman,E.P. *et al*. (2004) Expression profiling-best practices for data generation and interpretation in clinical trials. *Nat. Rev. Genet*., **5**, 229–237.

Junttila,M.R. and de Sauvage,F.J. (2013) Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature*, **501**, 346–354.

Moreau,E. (2001) A generalization of joint-diagonalization criteria for source separation. *IEEE Trans. Signal Process*., **49**, 530–541.

Yoshihara,K. *et al*. (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun*., **4**, 2612.

Yu,G. *et al*. (2011) BACOM: in silico detection of genomic deletion types and correction of normal cell contamination in copy number data. *Bioinformatics*, **27**, 1473–1480.