Bioinformatics, 34(11), 2018, 1969–1979
doi: 10.1093/bioinformatics/bty019
Advance Access Publication Date: 16 January 2018
Review

OXFORD

Gene expression

Computational deconvolution of transcriptomics data from mixed cell populations

Francisco Avila Cobos[1,2,3], Jo Vandesompele[1,2,3], Pieter Mestdagh[1,2,3,†] and Katleen De Preter[1,2,3,*,†]

[1]Center for Medical Genetics Ghent (CMGG), Ghent University, 9000 Ghent, Belgium, [2]Cancer Research Institute Ghent (CRIG), 9000 Ghent, Belgium and [3]Bioinformatics Institute Ghent from Nucleotides to Networks (BIG N2N), 9000 Ghent, Belgium

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.
Associate Editor: Jonathan Wren

Received on August 18, 2017; revised on October 31, 2017; editorial decision on January 2, 2018; accepted on January 10, 2018

Abstract

Summary: Gene expression analyses of bulk tissues often ignore cell type composition as an important confounding factor, resulting in a loss of signal from lowly abundant cell types. In this review, we highlight the importance and value of computational deconvolution methods to infer the abundance of different cell types and/or cell type-specific expression profiles in heterogeneous samples without performing physical cell sorting. We also explain the various deconvolution scenarios, the mathematical approaches used to solve them and the effect of data processing and different confounding factors on the accuracy of the deconvolution results.
Contact: katleen.depreter@ugent.be
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Over the years, the analysis of the transcriptome has substantially contributed to our understanding of the processes involved in human development and disease, but the complex nature of samples and tissues under investigation has been largely neglected.

For instance, most tumor samples are heterogeneous in nature, containing a variable portion of non-malignant cells that depends on the cancer type (Fridman *et al.*, 2012) (even when collected from the same patient) and includes epithelial, stromal and infiltrating immune cells (Egeblad *et al.*, 2010). Moreover, the expression level of each individual gene varies between different cell types and, when analyzing bulk samples of heterogeneous tissues, only tissue-averaged expression levels are measured. As a result, the expression contribution of low abundant cell types could be masked by that of more abundant ones (Kuhn *et al.*, 2012). Therefore, observed changes in gene expression might be the result of underlying differences in cell type proportions between samples, genuine changes due to clinical condition or a combination of both.

Since traditional analyses do not take into account cell type composition as a confounding factor in differential gene expression analyses, this might result in a loss of signal from less abundant cell types and might limit the conclusions that can be drawn from the experiments. Together with insufficiently documented or incorrect data processing practices (MAQC Consortium, 2010), platform-specific differences and variations introduced during the library construction (SEQC/MAQC-III Consortium, 2014), this sample heterogeneity may partially explain the problem of lack of reproducibility that the scientific community is currently facing (Baker, 2016).

The field of single-cell genomics has grown exponentially during the past few years, leading to the development of novel tools for the analysis of single cells within heterogeneous tissues (https://github.com/seandavi/awesome-single-cell). Initially, single cells were

isolated using Fluorescence-Activated Cell Sorting (FACS) or Laser Capture Microdissection (LCM) technologies. However, this up-front sorting required specific cell-surface markers to be known and appropriate antibodies to be available. To overcome this bottleneck, novel systems enable single cell isolation using other physical properties (e.g. cell size) by applying microfluidics or dielectrophoretic separation. Nevertheless, while promising, single-cell technologies have labour-intensive protocols and require expensive and specialized resources, currently hindering their establishment in a clinical setting.

For all these reasons, and to exploit the wealth of publicly available bulk data that can be re-analyzed, multiple computational approaches have been developed in the past years to infer abundance of different cell types and/or cell type-specific expression profiles in heterogeneous samples. This task, known as computational deconvolution of expression data from mixed cell populations, is the scope of this review. Therefore, articles regarding deconvolution of unrelated topics (i.e. fluorescence measurements or image processing) were not included here. The use of other omics data is briefly discussed in the section 'Alternative data types to perform the deconvolution'. Moreover Teschendorff and Zheng (2017) recently highlighted the impact of cell-type heterogeneity in DNA methylation data and provided a detailed overview of algorithms for correcting cell-type composition in the context of Illumina Infinium Methylation Beadarrays. Titus *et al.* (2017) have also recently published a review about cell-type deconvolution from DNA methylation.

To date, there are three reviews addressing the deconvolution problem using transcriptomics data as input (Mohammadi *et al.*, 2017; Shen-Orr and Gaujoux, 2013; Yadav and De, 2015; see Fig. 1). For those readers also interested in an in-depth analysis of the impact of different loss functions, constraint choices and regularization techniques on the deconvolution, we highly recommend the review by Mohammadi *et al.* (2017), where a total of 16 different combinations were carefully investigated.

Unlike the previous references, here we extensively review more than 50 methods developed between 2001 and 2017 (see Fig. 1; Supplementary Table S1). We also discuss the effect of data scale, (pre-)processing, normalization and different confounding factors on the accuracy of the deconvolution results. Finally, we highlight potential issues associated with traditional linear modelling and include a section with methods readily available as webtools, whose only requirement is an internet connection.

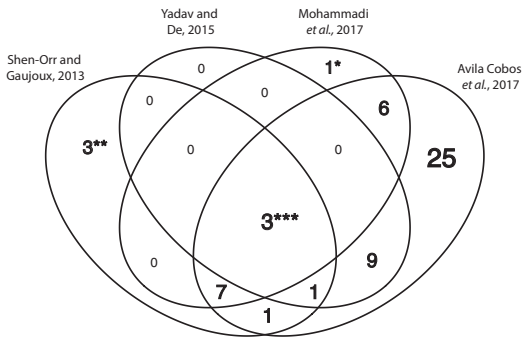

Fig. 1. Venn diagram showing the number of deconvolution methods covered by each review article (using transcriptomics data as input). (*) We discussed Nanodissection 1.0 within Section 4 as a method for selecting cell-type specific markers; (**) There are 3 methods only reviewed by Shen-Orr and Gaujoux (2013): Verhaak *et al.* (2010), Miller *et al.* (2011) and Shoemaker *et al.* (2012) [references taken from their article]; (***) There are three methods covered by all reviews: Erkkilä *et al.* (2010), Shen-Orr *et al.* (2010) and Qiao *et al.* (2012). Gosink et al., 2007; Bolen et al., 2011; Frishberg et al., 2016; Shen et al., 2016; Steuerman and Gat-Viks, 2016 are described on Supplementary Table S1

2 Defining the deconvolution problem

Sixteen years ago, Venet *et al.* (2001) presented a method '*to infer the gene expression profile of the various cellular types [...] directly from the measurements taken on the whole sample*'. They framed this problem as a multiple linear regression model applied to microarray gene expression data. The generalization of this problem belongs to the category of blind signal separation (BSS) problems, with the 'cocktail party problem' being the most known example (Cherry, 1953). It was formulated as the recognition of what a person says when others are speaking at the same time (Bronkhorst, 2015), which in turn can be translated into separating a set of observations into the constituent independent signals (sources).

The expression of a given gene in a heterogeneous sample can be modelled as the weighted sum (=linear combination) of the expression values from each cell type present in the mixture, assuming that every cell type has similar expression levels across different samples. Thus, the deconvolution problem can be formulated in matrix notation as follows (Equation 1):

$$T = C \cdot P \tag{1}$$

where T = measured expression values from heterogeneous tissue/tumor samples; C = cell type-specific average expression values and P = mixing proportions.

It can also be formulated algebraically as a latent variable model where the error term is not directly measurable (Equation 2):

$$t_{ij} = \sum_{k=1}^{K} c_{ik} \cdot p_{kj} + e_{ij};\ i = 1 \ldots M \text{ and } j = 1 \ldots N \tag{2}$$

where t_{ij} = observed expression value of gene i in heterogeneous sample j; c_{ik} = averaged expression value of gene i in cell type k; p_{kj} = proportion of cell type k in sample j; e_{ij} = error term; K = number of cell types; M = number of genes and N = number of samples.

The deconvolution can be performed if the system of linear equations has solution [the number of solutions of a system of linear equations can be determined using the Capelli-Fontené-Frobenius-Kronecker-Rouché theorem (Gorodentsev, 2016)]. Depending on the information used as input, the deconvolution can have several definitions, as described in Figure 2.

Finally, some methods model the heterogeneous samples as a two-component system (e.g. tumor and non-tumor) (Ahn *et al.*, 2013; Ghosh, 2004; Lähdesmäki *et al.*, 2005; Li and Xie, 2013; Wang *et al.*, 2015), whereas others increase the complexity by including three or more cell types in the mixture (Becht *et al.*, 2016; Hoffmann *et al.*, 2006; Li *et al.*, 2016; Stuart *et al.*, 2004; Wang *et al.*, 2010), getting as far as 22 (Newman *et al.*, 2015) or 25 different cell types (Chen *et al.*, 2017).

3 Mathematical approaches to solve the deconvolution problem

A detailed description of the particular deconvolution problem solved by each method, the necessary input data and their availability can be found in Supplementary Table S1.

The most commonly used group of methods is called ordinary least squares (OLS), linear least squares (LLS) or simply least

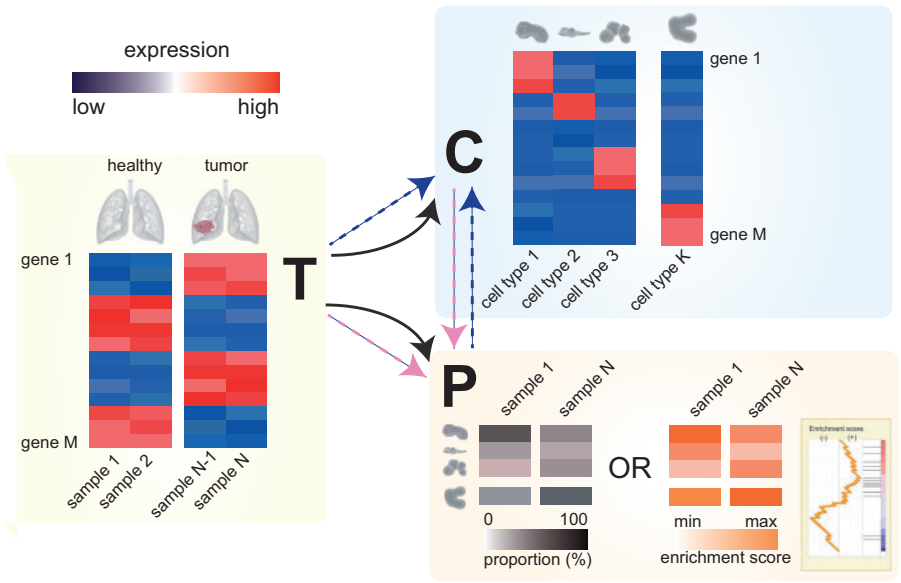

Fig. 2. The deconvolution problem has multiple formulations depending on the available input data. T = matrix containing the observed (measured) expression values from heterogeneous (tissue/tumor) samples (M genes, N samples); C = matrix consisting of cell type-specific average expression values (M genes, K cell types); P = matrix containing the mixing proportions (=relative composition) (K cell types, N samples); min = minimum; max = maximum. **Case 1)** Only T is available, C and P are estimated (dark grey arrows). **Case 2)** Given T and C, P is estimated (dashed pink arrows; grey heatmap on the bottom-right corner). One variant of this formulation uses T and cell type signatures (lists of marker genes for each cell type) known from literature or obtained by supervised/unsupervised marker selection strategies to estimate relative measures of the tissue heterogeneity (=enrichment scores; orange heatmap on the bottom-right corner) instead of cell type proportions [e.g. ESTIMATE (Yoshihara *et al.*, 2013), Şenbabaoğlu *et al.*, 2016]. Proportion values are strictly positive, bounded between 0 and 100 and with straightforward interpretation, whereas enrichment scores are unbounded and sometimes negative, making them harder to interpret. **Case 3)** Given T and P, C is estimated (dashed blue arrows). See Supplementary Table S1 and 'Mathematical approaches to solve the deconvolution problem' for more details

squares (LS), whose goal is to **minimize the sum of squares** of the differences between fitted (C·P) and observed values (T) (also known as minimization of the norm of the reconstruction error or minimization of the Euclidean distance) regardless of the distribution of the error term (Equation 3; see Fig. 2):

$$\text{Given T and C (or T and P)}: \ \min_{P(\text{or } C)} ||C \cdot P - T||^2 \quad (3)$$

Under the assumption that the error terms follow a normal distribution, a maximum likelihood estimation approach can also be applied to solve the minimization problem (Berkson, 1956).

Optimization problems aim to minimize or maximize diverse objective functions with or without imposed constraints. With this initial formulation of the problem (Equation 3, unconstrained optimization problem), both positive or negative proportions (P) may arise and the sum of the proportions might be different than one (see Box 1). Since those scenarios are meaningless in the context of the deconvolution, two constraints are included into the optimization problem: 1) the proportions must be strictly positive between 0 and 1 ('non-negativity' constraint); 2) the sum of proportions within each sample is 1 ('sum-to-one' constraint). This approach is known as the non-negative least squares method (NNLS) (Abbas *et al.*, 2009; Repsilber *et al.*, 2010; Venet *et al.*, 2001; Wang *et al.*, 2016; Zuckerman *et al.*, 2013). The nnls (https://nl.mathworks.com/matlabcentral/fileexchange/38003-nnls-non-negative-least-squares) or lsqnonneg (https://nl.mathworks.com/help/matlab/ref/lsqnonneg.html) functions in MATLAB or the nnls package (Stokkum and van, 2012) in R are common functions implementing this approach.

The sum of squared residuals can also be minimized using simulated annealing (SA) (Lu *et al.*, 2003; Wang *et al.*, 2006) (see Box 2) or other non-convex optimization strategies. However, since only convex objective functions guarantee that a local solution

Box 1. Dummy example for deconvolving cell type proportions of 4 cell types (k = 4) present in 1 sample (j = 1) assuming expression values in linear scale for 8 genes (i = 8).

Assuming T and C are known (P is unknown): Each cell type proportion corresponds to the regression coefficient (β) of a linear model formulated as:

$$t_{ij} = c_{ik} * \beta_{kj}$$

$$\begin{pmatrix} 1 \\ 10 \\ 850 \\ 1000 \\ 50 \\ 6 \\ 1080 \\ 1300 \end{pmatrix} = \begin{pmatrix} 20000 & 1 & 1 & 1 \\ 10000 & 1 & 1 & 1 \\ 1 & 1000 & 1 & 1 \\ 1 & 1000 & 1 & 1 \\ 1000 & 1 & 1000 & 1 \\ 20000 & 1 & 1000 & 1 \\ 1 & 1000 & 1 & 1000 \\ 1 & 1000 & 1 & 500 \end{pmatrix} * \begin{pmatrix} \beta_{1,1} \\ \beta_{2,1} \\ \beta_{3,1} \\ \beta_{4,1} \end{pmatrix}$$

When solving the above problem by linear least squares regression, the solution is: $\beta_{1,1} = -0.0005$, $\beta_{2,1} = 0.9789$, $\beta_{3,1} = 0.0320$ and $\beta_{4,1} = 0.2092$; with the total sum of proportions being 1.220.

Negative proportions are meaningless in the context of the deconvolution. When adding the *non-negativity* constraint by using the nnls function (R package), the new solutions are: $\beta_{1,1} = 0$, $\beta_{2,1} = 0.9789$, $\beta_{3,1} = 0.0268$ and $\beta_{4,1} = 0.2092$; with the total sum of proportions being 1.215. Finally, the *sum-to-one* constraint still has to be incorporated (during or after the optimization procedure) to obtain a definitive solution: $\beta_{1,1} = 0$, $\beta_{2,1} = 0.8057$, $\beta_{3,1} = 0.0221$ and $\beta_{4,1} = 0.1722$.

Box 2. Glossary of terms

Bayesian framework: statistical inference framework in which Bayes' theorem is used:

$$p(y|\theta) = \frac{p(\theta|y) * p(y)}{p(\theta)}$$

Therefore: $p(y|\theta) \propto p(\theta|y) * p(y)$ (where $\propto$ denotes proportionality). This is often translated into *posterior* $\propto$ *likelihood* $*$ *prior*

In Bayesian inference, the **prior distribution** represents the knowledge we have about how the data was generated before its actual generation. The prior is combined with the probability distribution of the observed data to yield the **posterior distribution**. The **likelihood** function for the data represents how likely the data (y) is given the model specified by any value of θ. A **parameter** is the numerical characteristic of a statistical model and a **hyperparameter** is the parameter of a prior distribution.

Condition number (CN) of a matrix: $||A|| * ||A^{-1}||$; where $||.||$ is the matrix norm. Example (using A and the Frobenius matrix norm):

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, A^{-1} = \begin{pmatrix} \frac{-1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{-1}{3} \end{pmatrix}$$

$$cond(A)_F = \sqrt{1^2 + 2^2 + 2^2 + 1^2} * \sqrt{\left(\tfrac{-1}{3}\right)^2 + \left(\tfrac{2}{3}\right)^2 + \left(\tfrac{2}{3}\right)^2 + \left(\tfrac{-1}{3}\right)^2} = 3.33$$

Convergence: criterion used to evaluate the improvement of a solution found by an algorithm after each iteration. When a solution does not change more than a pre-specified threshold with respect to the last n iterations ($n >= 1$), it is said that the algorithm has converged and it halts.

Convex function: function in which the midpoint of any segment between two points of the graph of the function is located above the graph or on the graph itself.

Frobenius matrix norm (=Euclidean norm): $\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_{ij}^2}$. Square root of the sum of the squares of the elements of a m x n matrix.

Gini index: in the context of marker selection, measure ranging from 0 to 1 used for the identification of tissue-enriched genes. The closer to 1, the higher the likelihood of a gene being exclusively expressed in one tissue.

ICA (Independent Component Analysis): unsupervised statistical technique that identifies mutually independent non-Gaussian components (dimensions) that are latent in the data. In contrast to PCA (where the components are uncorrelated and ranked by the amount of variance they explain), ICA components might be correlated.

Jensen-Shannon divergence: metric from information theory often used to discover cell-type/tissue specific genes. It quantifies the similarity between the expression of a given gene across cell types/tissues and that of a hypothetical gene whose expression is restricted to only one cell type/tissue.

K-dimensional polytope: geometric object of K dimensions with K flat sides.

L2-norm function: $\sqrt{\sum_{j=1}^{p}\beta_j^2}$. In the context of the deconvolution, β_j can be the least squares coefficient estimate and p the number of predictors.

Quadratic programming (QP): optimization of a function that contains at least one quadratic term and all constraints are linear.

Simulated Annealing (SA): optimization of a function that allows worse solutions at some iterations with a probability that decreases as the solution space is explored. The worsening steps allow a broader search across the function domain.

Sparse matrix: matrix in which most elements are zero.

Support Vector Regression (Support Vector Machine): supervised learning model used for regression or classification of linearly separable data into two categories.

corresponds to the global solution (Belzer *et al.*, 1979), these are not guaranteed to find the optimal solutions and might get stuck in local minima or maxima. Moreover, since gene expression matrices are typically not sparse, non-convex optimization strategies might result in high computational times and low rates of convergence (see Box 2).

Fortunately, every constrained LS problem will always be convex and a specific instance of a quadratic programming problem (Boyd and Vandenberghe, 2004) (see Box 2). Authors such as Gong *et al.* (2011) developed a method that guarantees a globally optimal solution by quadratic programming using the lsqlin function from MATLAB (https://nl.mathworks.com/help/optim/ug/lsqlin.html). Other commonly used functions are quadprog in MATLAB (Turlach and Weingessel, 2013; https://nl.mathworks.com/help/optim/ug/quadprog.html) and limSolve (Soetaert *et al.*, 2009; Shannon *et al.*, 2014) or the quadprog package (Turlach and Weingessel, 2013; Zhong *et al.*, 2013) in R Core Team (2017).

A second group of methods are **support vector regression approaches** with linear kernel (ν-SVR) (see Box 2), including CIBERSORT (Newman *et al.*, 2015) and ImmuCC (Chen *et al.*, 2017). Support vectors are robust against noise introduced by unknown cell types present in the mixture and involve the minimization of both a linear loss function and a L2-norm function (see Box 2 and '*Multicollinearity: presence of correlated cell types in the mixture*').

Since the first and second group of methods use *a priori* information (matrix C or P) along with T, they are often called supervised or guided approaches.

Thirdly, the separation of heterogeneous samples into their constituent cell types can be also approached as an unsupervised (=non-guided) **dimensionality reduction problem**, with principal component analysis (PCA) being widely used for this goal (Kassambara *et al.*, 2015; Lenz *et al.*, 2016). The number of relevant components (=cell types present) can be established visually or by using diverse rules (Peres-Neto *et al.*, 2005). However, PCA-based approaches may not be the most appropriate since factors other than the cell type identity might be contributing to the proportion of variance explained.

The final group of methods **jointly estimate** the gene expression of pure populations (C) and the mixing percentages (P) only starting from the expression data (T), without any prior information (=unsupervised or non-guided approaches; Case 1 from Fig. 1). It includes unsupervised non-negative matrix factorization (NMF or NNMF) and different Bayesian approaches. The **NMF formulation**

factorizes T as the product of C and P and incorporates the non-negativity constraint for all elements of both C and P. As a first step, initial values for P or C have to be generated (Gaujoux and Seoighe, 2012). On one hand, these initial values can be easily implemented by sampling random numbers from a uniform distribution. However, multiple attempts with different initializations are needed to achieve a stable final solution. (Moffitt *et al.*, 2015) successfully applied the NMF to pancreatic ductal adenocarcinoma with 20 random initializations, identifying different tumor subtypes with different tumor and stromal fractions. On the other hand, since the initialization process has a significant impact in the final results, singular value decomposition-based methods have been developed in an attempt to improve the initialization stage (Boutsidis and Gallopoulos, 2008). The most common algorithm used for NMF is called alternating least squares (ALS) (Berry *et al.*, 2007) and consists in two iterative steps that are repeated until convergence: first, P is fixed and, together with T, C is estimated by NNLS; secondly, C is fixed and, together with T, P is estimated.

Regarding those with a **Bayesian framework** (see Box 2), all attempt to maximize a likelihood function, but each method models the problem differently. They have a different type and number of parameters and hyper-parameters, with completely different *a priori* and *a posteriori* specifications (probability mass/density functions), leading to completely different likelihood functions. Since the joint estimation can be computationally intractable when the number of parameters is high, each method proposes different alternatives to make the problem tractable (e.g. approximating posterior distributions using Markov Chain Monte Carlo techniques, approximating expected values of parameters with Monte Carlo integration or using expectation-maximization (EM) algorithms to iteratively maximize the likelihood of the observed data) (Erkkilä *et al.*, 2010; Ghosh, 2004; Lähdesmäki *et al.*, 2005; Li and Xie, 2013; Roy *et al.*, 2006). It is unfeasible to describe them individually here and we highly advise the reader to go to the original publications to get a detailed overview of the modelling approach of interest.

The performance of non-guided methods strongly depends on the ability to recover meaningful gene signatures or expression profiles for the different cell types. Hence, supervised (=guided) methods where either C or P is available in addition to T, generally result in lower mean absolute difference (MAD) values (Gaujoux and Seoighe, 2012) and better performance (Gong *et al.*, 2011) compared to non-guided versions.

4 Selection of cell-type specific markers or expression profiles

For the second formulation of the deconvolution problem (see Fig. 2), cell type-specific markers or cell-type specific expression profiles are needed. In this section, we describe several useful approaches for this endeavour.

Importantly, we have focused on expression values at gene level throughout this review. However, scenarios where underlying differences in alternative transcript expression among different samples are masked at gene level may very well arise. Moreover, the usage of transcript expression might result in more candidate biomarkers and even higher cell-type specificities, potentially increasing the accuracy of the deconvolution results. Therefore, a deconvolution using expression values at both gene and transcript level should be considered if possible.

Ideally, a cell type-specific marker is a gene whose expression is restricted to one cell type and is robustly expressed across different biological replicates from the same cell type (Hoffmann *et al.*, 2006). However, since the deconvolution can only be solved if the number of marker genes is greater or equal than the number of cell types present in the mixture (Gorodentsev, 2016) and the presence of closely related cell types (=with only subtle differences in their transcriptome) is a very frequent scenario, the original restrictive definition of a cell type-specific marker is changed to a gene predominantly expressed in one cell type and to a lesser extent expressed in others (Venet *et al.*, 2001).

A first approach to select marker genes consists of finding genes whose average expression value in one cell type is several times greater than the median expression value across all cell types. The 'highly expressed, cell specific' (HECS) gene database (available at http://www.influenza-x.org/~jshoemaker/cten/f/HECS%20database.zip) (Shoemaker *et al.*, 2012) is an example of this approach and contains lists of cell type-specific genes from microarray data across 84 human and 96 murine tissues and cell types. The previous approach can be refined by statistically assessing differential gene expression between every cell type against all other cell types and setting arbitrary fold-change (e.g. ≥ 3) and p-value (e.g. <0.05) thresholds (Chen *et al.*, 2017; Chikina *et al.*, 2015; Gaujoux and Seoighe, 2012; Reinartz *et al.*, 2016). Of note, several authors recommend the use of medium-to-high expressed genes as robust markers, instead of the most expressed ones (Kuhn *et al.*, 2012; Repsilber *et al.*, 2010).

Some methods go one step further and rank the markers based on signal-to-noise ratios (Becht *et al.*, 2016; Wang *et al.*, 2006) or include an extra feature selection strategy to remove poorly discriminating marker genes (Kuhn *et al.*, 2012; Newman *et al.*, 2015; Shannon *et al.*, 2014). The F-statistic (measure of their fit in the multiple linear regression model) (Wang *et al.*, 2010), the Gini index (Zhang *et al.*, 2017), the Jensen-Shannon divergence (Cabili *et al.*, 2011) (see Box 2) or the components from PCA, ICA or NMF analyses (Zinovyev *et al.*, 2013) can be also used to identify marker genes.

More advanced methodologies include CellMapper (Nelms *et al.*, 2016), Nanodissection 1.0 (Ju *et al.*, 2013), UNDO (Wang *et al.*, 2015) and CAM (Wang *et al.*, 2016). Assuming that marker genes for a given cell type should correlate with each other and starting with as little as one cell type-specific marker gene, CellMapper (developed and validated using microarray data but potentially applicable to RNA-seq data) uses thousands of publicly available expression profile datasets (pre-loaded as objects in 'CellMapperData' Bioconductor package) or custom datasets to find other marker genes with similar expression patterns and specifically expressed in a cell type of interest.

By selecting among 28 different human tissues and uploading a set of at least ten candidate marker genes ('positive standard') and ten genes expressed in other lineages ('negative standard'), Nanodissection 1.0 (available at http://nano.princeton.edu/) estimates the probability that a gene is cell-type specific using an iterative linear support vector machine (SVM) approach.

Both UNDO and CAM are completely unsupervised approaches that allow novel marker identification without any prior information by geometrically identifying the vertices and resident genes of a K-dimensional polytope (see Box 2) built from a gene expression matrix, with K being the number of cell types present in the mixture.

In conclusion, the generation of cell type-specific expression matrices is not trivial, varies from method to method and is a determinant factor to consider when approaching the deconvolution strategy.

5 Factors affecting the deconvolution efficiency

Several studies have shown that the detection of differentially expressed genes after the deconvolution of bulk expression data is less

prone to the identification of false positives and false negatives (Wang *et al.*, 2006), resulting in more accurate (Zhong *et al.*, 2013), specific and sensitive results (Shen-Orr *et al.*, 2010) when compared to those obtained from bulk heterogeneous (tumor) samples.

Cancer genomics is a field that can greatly benefit from the deconvolution of bulk expression data. Taking tumor heterogeneity into account led to an increase in the sensitivity of relapse-free survival analyses and more accurate tumor subtype predictions (Elloumi *et al.*, 2011). Nevertheless, it is important to note that tumor heterogeneity is challenging at multiple levels: first, inter-tumor heterogeneity exist both between different tumor types and between samples within a given cancer (sub-)type (biological heterogeneity). Second, intra-tumor heterogeneity may also exist within a given sample (different tumor subclones). Several authors stated that their methods should specifically be applied to samples belonging to a common tumor (sub-)type (Ahn *et al.*, 2013; Quon and Morris, 2009) or to a common tissue (Frishberg *et al.*, 2015). Importantly, non-guided approaches such as NMF successfully identified different pancreatic ductal adenocarcinoma subtypes (Moffitt *et al.*, 2015). However, this only addresses the inter-tumor heterogeneity. In order to study intra-tumor heterogeneity, either single-cell profiling data or sequencing multiple locations from the same tumor would be needed.

Of note, there are multiple factors affecting the performance of the deconvolution, which are discussed below.

5.1 Effect of pre-processing and normalization

As Hoffmann *et al.* (2006), Clarke *et al.* (2010) and Repsilber *et al.* (2010) pointed out, the data normalization procedure has an impact on the estimation of cell type proportions, cell type-specific expression profiles and thus, the power to detect differential expression. Moreover, (Newman *et al.*, 2017) highlighted the need of accounting for normalization differences in order to perform meaningful comparisons between different deconvolution methods. Most methods presented in this review assume that the data is appropriately pre-processed and normalized prior to the deconvolution (see Supplementary Table S1). Some methods applied to data coming from different platforms incorporate a batch effect correction using Combat (Gentles *et al.*, 2015; Şenbabaoğlu *et al.*, 2016) or the supervised normalization of microarray (SNM) method (Qiao *et al.*, 2012).

Controversially, some methods propose background correction (Chen *et al.*, 2017; Shannon *et al.*, 2014) whereas others recommend not to apply it (Liebner *et al.*, 2014). On the one hand, (Hoffmann *et al.*, 2006) finds the Microarray Suite 5 (MAS5) to provide a more robust estimation of the proportions compared to the robust multi-array analysis (RMA) and model based expression index (MBEI). On the other hand, (Ahn *et al.*, 2013) discuss a deviation from the linearity assumption when applying MAS5 scale normalization, which was not observed when using RMA together with quantile normalization. Interestingly, (Irizarry *et al.*, 2006) shows that rather than the normalization method, background correction is the main factor explaining differences between different pre-processing alternatives for Affymetrix GeneChip systems. Thus, a quantitative evaluation of the impact on the deconvolution results would be relevant for the field but is outside the scope of this review. A detailed summary about the normalization strategies can be found in Supplementary Table S1.

5.2 Logarithmic versus linear space

Statistical tests typically used to assess differential gene expression assume an underlying normal distribution of the data being analyzed. For this reason, since the log-normal distribution is considered as a good approximation for microarray expression data (Hoyle *et al.*, 2002) and stabilizes the variance (Tsai *et al.*, 2003), the data is often transformed into logarithmic scale.

However, Zhong and Liu (2012) showed that log transformed microarray data violated the linearity assumption of Equation 2 (see 'Defining the deconvolution problem'), leading to a consistent under-estimation of the signal when deconvolving cell-type specific expression profiles. On the other hand, when the data was transformed back to linear scale, it resulted in an accurate deconvolution. In our experience, data transformations used for variance stabilization (e.g. glog transformation) also lead to a consistent under-estimation, in agreement with Zhong and Liu (2012) (data not shown). The linearity assumption was also confirmed by Kuhn *et al.* (2011, 2012) on non-log transformed microarray data. Zhong *et al.* (2013) alleged that the linearity assumption also holds true for RNA-seq data and recently, Jin *et al.* (2017) performed a thorough assessment of the linearity assumption of transcript abundance from RNA-seq data. They showed the need of normalizing the data prior to the deconvolution and concluded that when using RNA-seq data, TPM values from Salmon, RSEM or Kallisto provided the most accurate reconstruction of cell type proportions present in a mixture.

In line with this argument, the vast majority of methods reviewed here agreed on transforming the data into log scale for pre-processing and data normalization followed by a conversion back to linear scale (using the anti-log transformation) prior to the deconvolution (Anghel *et al.*, 2015; Lähdesmäki *et al.*, 2005; Wang *et al.*, 2016). Although the linearity assumption is valid for most genes, a more accurate deconvolution might be achieved by detecting and excluding genes affected by non-linear amplification (Shen-Orr *et al.*, 2010), excluding noisy genes with little biological signal (Abbas *et al.*, 2009) or removing outliers (=trimmed robust regression) (Hoffmann *et al.*, 2006) before applying the least squares method.

However, others claimed that it is possible to apply the deconvolution to both log-transformed and non-log transformed data (Erkkilä *et al.*, 2010; Repsilber *et al.*, 2010), modelled the expression data as log2 normal distributions (Clarke *et al.*, 2010; Elloumi *et al.*, 2011) or claimed more accurate results when using quantile normalized and log2-transformed data (Shannon *et al.*, 2014). Furthermore, Clarke *et al.* (2010) requires log-transformed data to find accurate estimates of the proportion of a cell type in a mixture.

A counterintuitive statement comes from Repsilber *et al.* (2010), claiming optimal deconvolution of cell type-specific gene expression using log-transformed data whereas cell type-specific differential expression is optimal when using non-log-transformed data.

Since the cell type-specific expression is over-estimated when the linear relationship between the heterogeneous expression and the cell proportion predictors is absent and the interpretation of the regression coefficients may be incorrect, we advise to perform the deconvolution on data in linear scale.

5.3 Multicollinearity: presence of correlated cell types in the mixture

Significant correlation between two or more cell types (also known as multicollinearity in the context of linear regression) might result in an increase of the estimation errors and the impossibility of separating the contribution from individual cell types (Kuhn *et al.*, 2012).

Even though some authors assume gene expression profiles between different cell fractions to be uncorrelated (Venet *et al.*, 2001), this might be an unrealistic scenario with important consequences.

As (Newman *et al.*, 2015) pointed out, the deconvolution results can be negatively affected when many related cell types were present, which may result in higher proportions being assigned to the cell type whose expression profile is most similar to the mixture. One possible solution to tackle this problem is the support vector regression (SVR) methodology implemented by CIBERSORT (Newman *et al.*, 2015), which minimizes the variance of weights assigned to highly correlated predictors. CIBERSORT was able to deal with five highly collinear cell types and has been successfully applied to more than 18 000 expression profiles to analyze overall survival across 25 cancer types and abundance of diverse tumor-associated leukocyte subsets (Gentles *et al.*, 2015).

Mohammadi *et al.* (2017) found that using the L2 loss function together with an R2 regularizer gave the best results and they reasoned that the regularization of the objective function can improve the performance in cases where highly correlated cell types are present in a mixture.

5.4 Condition number of a matrix

It is known that the condition number ('CN'; see Box 2) has an impact when solving systems of linear equations (Equation 1) (Fang, 2003). Abbas *et al.* (2009) and Newman *et al.* (2015) stated that reference expression profiles (matrix C in Fig. 2) could become more robust by minimizing the CN. Abbas *et al.* (2009) found the CN to be high for matrices containing small or large number of genes whereas the CN was minimum for moderate numbers. Newman *et al.* (2015) calculated the CN value for all candidate signature matrices for 22 cell types and kept the one with lowest CN. Glass and Dozmorov (2016) discovered that a high CN of the matrix containing the cell proportions (matrix P in Fig. 2) negatively affected the sensitivity of the deconvolution. Interestingly, Gentles *et al.* (2015) noticed that the exclusion of cell types with the lowest proportion mean resulted in a noticeable improvement in sensitivity and in a considerable reduction of the CN. Interestingly, Teschendorff and Zheng (2017) also emphasize the importance of optimizing the CN when selecting CpGs to deconvolute DNA methylation data. For all these reasons, we advise users to not overlook this factor when building the necessary matrices for their deconvolution problem, aiming at the smallest CN values as possible.

5.5 Cell cycle

Cells are dynamic systems, reflected by continuous changes in their transcriptome. Each sample has a mixture of cells in different phases of the cell cycle. When working with cultured cells, the cell cycle can be synchronized by chemical arrest or nutrient starvation (Bar-Joseph *et al.*, 2008). However, this is not possible when tissue samples are profiled. Lu *et al.* (2003) pioneered the estimation of the proportions of cells in different phases of the cell cycle using microarray expression data. They proposed the use of phase-specific markers (such as cyclin *CLN2* for phase G1 or *CLB4* for phase G2) to establish different time points of the cell cycle. Even though the vast majority of methods in the present review did not include this complex aspect when modelling the deconvolution problem, this must be ideally taken into account when developing new tools.

6 Minimum cell type proportions that can be detected

Zhong *et al.* (2013) were able to accurately estimate cell types present at more than 10%, with a substantial decrease in accuracy if the percentage was smaller than that threshold. PERT (Qiao *et al.*, 2012) and DeconRNAseq (Gong and Szustakowski, 2013) were able to retrieve proportions as small as 2% whereas CIBERSORT (Newman *et al.*, 2015) detected fractions down to 0.5% in mixtures containing $<50\%$ of tumor content.

7 Assessment of the deconvolution results

Multiple empirical approaches have been proposed to assess the validity of the estimations generated by the deconvolution methods: (i) in-situ hybridization (ISH) (Kuhn *et al.*, 2011, 2012) or immunohistochemistry (IHC) staining from the Human Protein Atlas (Ju *et al.*, 2013) to validate cell type-specific gene expression; (ii) comparison of predicted proportions with those measured by flow cytometry (Qiao *et al.*, 2012); (iii) combination of microscopy and FACS analysis to evaluate the estimated proportion of yeast cells in different stages of the cell cycle (Wang *et al.*, 2016); (iv) correlation with immune-fluorescence cell estimates or cell fractions inferred from DNA methylation (Li *et al.*, 2016; Şenbabaoğlu *et al.*, 2016) or DNA copy number data (Şenbabaoğlu *et al.*, 2016).

8 Potential issues with traditional linear modelling

There are four important aspects that need to be taken into account when modelling gene expression data as the weighted sum of gene expression profiles of pure populations:

1) **There should be reference profiles for all populations present in the mixture or at least one marker for each cell type**. This might be problematic for some cell types that cannot be isolated easily (mostly the less abundant ones) and might not have been analyzed or sequenced yet. Since reference profiles are assumed to accurately represent the actual cell types present in heterogeneous samples (Qiao *et al.*, 2012), they should be carefully obtained. Moreover, the existence of a sufficient number of marker genes to perform the deconvolution is crucial (Hoffmann *et al.*, 2006). Some methods need as little as one marker per cell type (Venet *et al.*, 2001) but most of them recommend a higher number (5–10) to avoid the potential influence of outliers (Ahn *et al.*, 2013).

2) **Since the true composition is unknown, some cell types may be ignored.** Some methods require precise knowledge of either the constituent cell types (Kuhn *et al.*, 2012) or the cell type proportions present in the heterogeneous sample (Li and Xie, 2013) (e.g. assessment from a pathologist or estimated by FACS) for solving the deconvolution problem. However, it is possible that there are no surface markers available yet (Altboum *et al.*, 2014) for sorting unknown populations. Moreover, since the assessment of a pathologist provides information about cell type proportions but not on the amount of mRNA present, the estimates might not be accurate (Clarke *et al.*, 2010). Even though we have stated that some unsupervised methods take advantage of *a priori* information whenever this is available, other authors are against this practice. For example, (Chikina *et al.*, 2015) argue that Coulter counter measurements can have an error $\geq 5\%$ for lowly abundant cell types, advising not to use them as input for the deconvolution. Furthermore, Gong *et al.* (2011) showed that Erkkila's Bayesian approach could not find any solutions when seeded with random estimates (=absence of prior information). Therefore, although *a priori* information can be efficiently exploited (e.g. in a Bayesian framework), the use of incorrect proportion estimates can negatively affect the deconvolution. Finally, an incorrect model specification (e.g. ignoring a cell type that is actually present) might result in incorrect estimates

of cell type-specific expression levels for some methods (Kuhn *et al.*, 2012; Zuckerman *et al.*, 2013).

3) **Some methods designed to infer the cell type composition from expression data assume a stable cell type composition within a given heterogeneous tissue** (Ahn *et al.*, 2013). Marker genes are not guaranteed to be expressed at the same levels across different cells (Zhong *et al.*, 2013), even in a tumor from the same patient. Furthermore, the expression profiles are platform-specific, which might result in markers not being present in all platforms and in varying expression levels for a given marker across different platforms (Li and Xie, 2013; Shannon *et al.*, 2017).

Assuming that the expression of a marker gene in one cell type is independent from other cell types present in the mixture is often unrealistic due to potential paracrine signalling effects. This can be tackled by including an extra coefficient in the linear model accounting for the cross-product between different cell types: Kuhn *et al.* (2012) excluded all those genes likely to be expressed by a cell type that was not included in the model and Stuart *et al.* (2004) observed many transcripts with high cross-product values, suggesting that the expression levels in one cell type are affected by the presence and abundance of other cell types.

4) **The majority of the methods do not take into account the fact that the reference expression profiles are often perturbed by microenvironment or developmental effects or were simply obtained under different conditions or with different technologies or platforms**. To address this issue, PERT (Qiao *et al.*, 2012) estimates a shared perturbation factor across all cell types to account for transcriptional variation between the reference and constituent expression profiles. ISOLATE (Quon and Morris, 2009) uses a multinomial model to measure noise in gene expression data and assumes that there is a new population not represented by the available reference profiles. Finally, ISOpure (Quon *et al.*, 2013) [ISOpureR (Anghel *et al.*, 2015)] is similar to ISOLATE in the estimation of tumor purities and a reference cancer profile but assumes that each healthy profile is the weighted sum of the available healthy tissue profiles and imposes non-negative and sum-to-one constraints.

9 Deconvolution methods readily available as webtools

The column 'Availability/GUI' from Supplementary Table S1 contains detailed information about how to get access to the different reviewed methods. Most of them are accessible as pre-built packages or raw code from different programming languages (e.g. R, Python, Java, ...). For scientists lacking bioinformatics skills, we highlight seven tools readily accessible for everyone with an internet connection, with little or no bioinformatics background required:

- CellPred (Wang *et al.*, 2010): Allows estimation of cell type proportions using Affymetrix microarray data as input. Available at http://webarraydb.org/webarray/index.html (CellPred tab).
- TIMER (Li *et al.*, 2016): A great resource containing the proportions of B cells, CD4+ and CD8+ T cells, macrophages, neutrophils and dendritic cells across 11 509 samples corresponding to 32 cancer types from The Cancer Genome Atlas (TCGA). Available at https://cistrome.shinyapps.io/timer/. Users can download the TIMER method from https://github.com/hanfeisun/TIMER to run it on their own samples.
- DSection (Erkkilä *et al.*, 2010): Estimation of cell type-specific expression profiles, corrected cell type proportions and differential gene expression using microarray data. Available at: http://informatics.systemsbiology.net/DSection/.
- DCQ (Altboum *et al.*, 2014) and CoD (Frishberg *et al.*, 2015) are two tools from the Irit Gat-Viks lab allowing the estimation of cell type quantities to identify disease-relevant cell types using microarray or RNA-seq data. Available at: http://www.dcq.tau.ac.il/ (detailed information: http://dcq.tau.ac.il/application.html) and http://www.csgi.tau.ac.il/CoD/ (detailed information: http://www.csgi.tau.ac.il/CoD/application.html).
- ESTIMATE (Yoshihara *et al.*, 2013): Allows quick access to relative stromal and immune cell type composition across all samples available at TCGA (microarray and RNA-seq data). Available at: http://bioinformatics.mdanderson.org/estimate/.
- CIBERSORT (Newman *et al.*, 2015): Given microarray or RNA-seq data from heterogeneous samples and selecting pre-built or custom-made matrices with cell type-specific expression profiles, it generates proportions of up to 22 cell types. Available at: https://cibersort.stanford.edu/runcibersort.php.

10 Alternative data types to perform the deconvolution

Although being outside of the scope of this review, other omics data also being used as input for the deconvolution problem are worth mentioning due to their rapid growth.

EpiDISH (Teschendorff *et al.*, 2017) infers cell-type composition using DNA methylation data and cell-type specific DNase hypersensitive sites. Other tools such as MeDeCom (Lutsik *et al.*, 2017) and eFORGE (Breeze *et al.*, 2016) have been designed to estimate cell type-specific signal and account for tumor purity in heterogeneous methylomes. Onuchic *et al.* (2016) proposed EDec, a two-step approach in which cell-type proportions in each sample and cell type-specific methylation and gene expression profiles are retrieved. Importantly, as Teschendorff and Zheng (2017) pointed out, a direct comparison between expression-based and DNA methylation-based cell type composition estimates has not been performed yet.

Several methods have been proposed to detect copy number aberrations from DNA profiling of heterogeneous samples: BACOM 2.0 (Fu *et al.*, 2015), ABSOLUTE (Carter *et al.*, 2012) and CloneCNA (Yu *et al.*, 2016). Finally, Aran *et al.* (2015) created the Consensus measurement of Purity Estimation (CPE), a robust value for tumor purity obtained from combining gene expression, somatic copy number, methylation and immunohistochemistry data that they successfully applied to more than 10 000 samples from The Cancer Genome Atlas (TCGA).

11 Conclusion and future directions

Bayesian and regression-based methodologies have been proven effective in the framework of the deconvolution problem. However, currently there is no tool addressing all the challenges we discussed throughout this review, leaving some room for improvement. The ideal tool should: (i) include alternatives to solve all formulations of the deconvolution problem described in Figure 2, meaning supervised and completely unsupervised scenarios. For the former scenario and following the concerns we raised 'Potential issues with traditional linear modelling', we argue against the use of non-informative (=random) initial estimates and recommend the use of one or more approaches described in '*Selection of cell type-specific markers or expression profiles*'. For the latter we propose the geometric identification of markers proposed by UNDO (Wang *et al.*, 2015)

and CAM (Wang *et al.*, 2016), as they only rely on the geometric topology inherent to the expression data from a mixture rather than external reference datasets (that might come from several technology platforms) or arbitrary log fold change and p-value thresholds; (ii) allow to study the changes in cell type proportions across multiple time points [such as DCQ (Altboum *et al.*, 2014)]; (iii) account for different phases of the cell cycle using markers such as CLN2 for phase G1; (iv) account for small perturbations between reference expression profiles of pure cell types and those constituting the heterogeneous samples [such as PERT (Qiao *et al.*, 2012) or ISOpure(R) (Quon *et al.*, 2013; Anghel *et al.*, 2015)]; (v) be computationally efficient, with fast running time and rate of convergence; (vi) be able to account for the presence of multiple correlated cell types in the mixture [such as CIBERSORT (Newman *et al.*, 2015)].

The amount of gene expression data from single cells is growing exponentially, revealing information that is hidden in tissue-averaged expression measurements from heterogeneous samples. However, the expression levels are often smaller than the detection limits of current state-of-the-art single-cell technologies. To overcome the detection issue, an approach called 'stochastic profiling' has been proposed (Bajikar *et al.*, 2014; Janes *et al.*, 2010; Narayanan *et al.*, 2016). Stochastic profiling consists of measuring the expression of random pools of cells (e.g. 10 cells) followed by modelling the expression of each gene as a binomial choice from a mixture of two different regulatory states: 'ON' for cells expressing the gene and 'OFF' for those that do not. Since the amount of input mRNA from a pool of cells is bigger than the mRNA from a single cell, this method offers more robust detection.

In conclusion, while single-cell and stochastic profiling are postulated as firm candidates to revolutionize the transcriptomics field with continuous improvements in terms of sensitivity and affordability, we foresee a rapid inclusion of deconvolution methodologies to existing pipelines for the analysis of omics data in the meantime, increasing the accuracy and reliability of downstream cell type-specific differential gene expression analysis without incurring in additional costs.

Funding

This work was supported by a Special Research Fund (BOF) scholarship of Ghent University to Francisco Avila Cobos (BOF.DOC.2017.0026.01).

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. *Bioinf. Oxf. Engl.*, **29**, 1865–1871.

Altboum,Z. *et al.* (2014) Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol. Syst. Biol.*, **10**, 720.

Anghel,C.V. *et al.* (2015) ISOpureR: an R implementation of a computational purification algorithm of mixed tumour profiles. *BMC Bioinformatics*, **16**, 156.

Aran,D. *et al.* (2015) Systematic pan-cancer analysis of tumour purity. *Nat. Commun.*, **6**, 8971.

Bajikar,S.S. *et al.* (2014) Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. *Proc. Natl. Acad. Sci. USA*, **111**, E626–E635.

Baker,M. (2016) 1,500 scientists lift the lid on reproducibility. *Nat. News*, **533**, 452.

Bar-Joseph,Z. *et al.* (2008) Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proc. Natl. Acad. Sci. USA*, **105**, 955–960.

Becht,E. *et al.* (2016) Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.*, **17**, 218.

Belzer,J. *et al.* (1979) *Encyclopedia of Computer Science and Technology: Volume 11 - Minicomputers to PASCAL*. CRC Press, New York, USA.

Berkson,J. (1956) *Estimation by Least Squares and by Maximum Likelihood*. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, 1–11, University of California Press, Berkeley, California.

Berry,M.W. *et al.* (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.*, **52**, 155–173.

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

Boutsidis,C. and Gallopoulos,E. (2008) SVD based initialization: a head start for nonnegative matrix factorization. *Pattern Recognit.*, **41**, 1350–1362.

Boyd,S. and Vandenberghe,L. (2004) *Convex Optimization*. Cambridge University Press, Cambridge, UK.

Breeze,C.E. *et al.* (2016) eFORGE: a tool for identifying cell type-specific signal in epigenomic data. *Cell Rep.*, **17**, 2137–2150.

Bronkhorst,A.W. (2015) The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Atten. Percept. Psychophys.*, **77**, 1465.

Cabili,M.N. *et al.* (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.

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

Chen,Z. *et al.* (2017) Inference of immune cell composition on the expression profiles of mouse tissue. *Sci. Rep.*, **7**, 40508.

Cherry,E.C. (1953) Some Experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.*, **25**, 975–979.

Chikina,M. *et al.* (2015) CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinf. Oxf. Engl.*, **31**, 1584–1591.

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

Egeblad,M. *et al.* (2010) Tumors as organs: complex tissues that interface with the entire organism. *Dev. Cell*, **18**, 884–901.

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.

Erkkilä,T. *et al.* (2010) Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinf. Oxf. Engl.*, **26**, 2571–2577.

Fang,Q. (2003) A note on the condition number of a matrix. *J. Comput. Appl. Math.*, **157**, 231–234.

Fridman,W.H. *et al.* (2012) The immune contexture in human tumours: impact on clinical outcome. *Nat. Rev. Cancer*, **12**, 298–306.

Frishberg,A. *et al.* (2015) CoD: inferring immune-cell quantities related to disease states. *Bioinf. Oxf. Engl.*, **31**, 3961–3969.

Frishberg,A. *et al.* (2016) ImmQuant: a user-friendly tool for inferring immune cell-type composition from gene-expression data. *Bioinf. Oxf. Engl.*, **32**, 3842–3843.

Fu,Y. *et al.* (2015) BACOM2.0 facilitates absolute normalization and quantification of somatic copy number alterations in heterogeneous tumor. *Sci. Rep.*, **5**, 13955.

Gaujoux,R. and Seoighe,C. (2012) Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.*, **12**, 913–921.

Gentles,A.J. *et al.* (2015) The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.*, **21**, 938–945.

Ghosh,D. (2004) Mixture models for assessing differential expression in complex tissues using microarray data. *Bioinf. Oxf. Engl.*, **20**, 1663–1669.

Glass,E.R. and Dozmorov,M.G. (2016) Improving sensitivity of linear regression-based cell type-specific differential expression deconvolution with per-gene vs. global significance threshold. *BMC Bioinformatics*, **17**, 334.

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. *Bioinf. Oxf. Engl.*, **29**, 1083–1085.

Gorodentsev,A.L. (2016) *Algebra I: Textbook for Students of Mathematics*. Springer, Cham, Switzerland.

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

Hoffmann,M. *et al.* (2006) Robust computational reconstitution – a new method for the comparative analysis of gene expression in tissues and isolated cell fractions. *BMC Bioinformatics*, **7**, 369.

Hoyle,D.C. *et al.* (2002) Making sense of microarray data distributions. *Bioinformatics*, **18**, 576–584.

Irizarry,R.A. *et al.* (2006) Comparison of affymetrix GeneChip expression measures. *Bioinformatics*, **22**, 789–794.

Janes,K.A. *et al.* (2010) Identifying single-cell molecular programs by stochastic profiling. *Nat. Methods*, **7**, 311–317.

Jin,H. *et al.* (2017) Comprehensive evaluation of RNA-seq quantification methods for linearity. *BMC Bioinformatics*, **18**, 117.

Ju,W. *et al.* (2013) Defining cell-type specificity at the transcriptional level in human disease. *Genome Res.*, **23**, 1862–1873.

Kassambara,A. *et al.* (2015) GenomicScape: an easy-to-use web tool for gene expression data analysis. Application to investigate the molecular events in the differentiation of B cells into plasma cells. *PLOS Comput. Biol.*, **11**, e1004077.

Kuhn,A. *et al.* (2012) Cell population-specific expression analysis of human cerebellum. *BMC Genomics*, **13**, 610.

Kuhn,A. *et al.* (2011) Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat. Methods*, **8**, 945–947.

Lähdesmäki,H. *et al.* (2005) In silico microdissection of microarray data from heterogeneous cell populations. *BMC Bioinformatics*, **6**, 54.

Lenz,M. *et al.* (2016) Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. *Sci. Rep.*, **6**, 25696.

Li,B. *et al.* (2016) Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.*, **17**, 174.

Li,Y. and Xie,X. (2013) A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. *BMC Bioinformatics*, **14**, S11.

Liebner,D.A. *et al.* (2014) MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinf. Oxf. Engl.*, **30**, 682–689.

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.

Lutsik,P. *et al.* (2017) MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. *Genome Biol.*, **18**, 55.

MAQC Consortium. (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.*, **28**, 827–838.

Miller,J.A. *et al.* (2011) Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinformatics*, **12**, 322.

Moffitt,R.A. *et al.* (2015) Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.*, **47**, 1168–1178.

Mohammadi,S. *et al.* (2017) A critical survey of deconvolution methods for separating cell types in complex tissues. *Proc. IEEE*, **105**, 340–366.

Narayanan,M. *et al.* (2016) Robust inference of cell-to-cell expression variations from single- and K-cell profiling. *PLOS Comput. Biol.*, **12**, e1005016.

Nelms,B.D. *et al.* (2016) CellMapper: rapid and accurate inference of gene expression in difficult-to-isolate cell types. *Genome Biol.*, **17**, 201.

Newman,A.M. *et al.* (2017) Data normalization considerations for digital tumor dissection. *Genome Biol.*, **18**, 128.

Newman,A.M. *et al.* (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, **12**, 453–457.

Onuchic,V. *et al.* (2016) Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types. *Cell Rep.*, **17**, 2075–2086.

Peres-Neto,P.R. *et al.* (2005) How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput. Stat. Data Anal.*, **49**, 974–997.

Qiao,W. *et al.* (2012) PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput. Biol.*, **8**, e1002838.

Quon,G. *et al.* (2013) Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med.*, **5**, 29.

Quon,G. and Morris,Q. (2009) ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinf. Oxf. Engl.*, **25**, 2882–2889.

R Core Team. (2017) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reinartz,S. *et al.* (2016) A transcriptome-based global map of signaling pathways in the ovarian cancer microenvironment associated with clinical outcome. *Genome Biol.*, **17**, 108.

Repsilber,D. *et al.* (2010) Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics*, **11**, 27.

Roy,S. *et al.* (2006) A hidden-state Markov model for cell population deconvolution. *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, **13**, 1749–1774.

Şenbabaoğlu,Y. *et al.* (2016) Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol.*, **17**, 231.

SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.*, **32**, 903–914.

Shannon,C.P. *et al.* (2017) Enumerateblood – an R package to estimate the cellular composition of whole blood from Affymetrix Gene ST gene expression profiles. *BMC Genomics*, **18**, 43.

Shannon,C.P. *et al.* (2014) Two-stage, in silico deconvolution of the lymphocyte compartment of the peripheral whole blood transcriptome in the context of acute kidney allograft rejection. *PloS One*, **9**, e95224.

Shen,Q. *et al.* (2016) contamDE: differential expression analysis of RNA-seq data for contaminated tumor samples. *Bioinf. Oxf. Engl.*, **32**, 705–712.

Shen-Orr,S.S. *et al.* (2010) Cell type-specific gene expression differences in complex tissues. *Nat. Methods*, **7**, 287–289.

Shen-Orr,S.S. and Gaujoux,R. (2013) Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.*, **25**, 571.

Shoemaker,J.E. *et al.* (2012) CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data. *BMC Genomics*, **13**, 460.

Soetaert,K. *et al.* (2009) limSolve: Solving Linear Inverse Models. R-package version 1.5.1, https://cran.r-project.org/web/packages/limSolve/citation.html.

Stokkum,K.M.M. and van,I.H.M. (2012) nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS).

Steuerman,Y. and Gat-Viks,I. (2016) Exploiting gene-expression deconvolution to probe the genetics of the immune system. *PLoS Comput. Biol.*, **12**, e1004856.

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.

Teschendorff,A.E. *et al.* (2017) A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics*, **18**,

Teschendorff,A.E. and Zheng,S.C. (2017) Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics*, **9**, 757–768.

Titus,A.J. *et al.* (2017) Cell-type deconvolution from DNA methylation: a review of recent applications. *Hum. Mol. Genet.*, **26**, R216–R224.

Tsai,C.-A. *et al.* (2003) Testing for differentially expressed genes with microarray data. *Nucleic Acids Res.*, **31**, e52.

Turlach,B.A. and Weingessel,A. (2013) *quadprog: Functions to solve Quadratic Programming Problems*. R package version 1.5–5, https://CRAN.R-project.org/package=quadprog.

Venet,D. *et al.* (2001) Separation of samples into their constituents using gene expression data. *Bioinformatics*, **17**, S279–S287.

Verhaak,R.G.W. et al. (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17, 98–110.

Wang,M. *et al*. (2006) Computational expression deconvolution in a complex mammalian organ. *BMC Bioinformatics*, **7**, 328.

Wang,N. *et al*. (2016) Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci. Rep*., **6**, 18909.

Wang,N. *et al*. (2015) UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinf. Oxf. Engl*., **31**, 137–139.

Wang,Y. *et al*. (2010) In silico estimates of tissue components in surgical samples based on expression profiling data. *Cancer Res*., **70**, 6448–6455.

Yadav,V.K. and De,S. (2015) An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Brief. Bioinf*., **16**, 232–241.

Yoshihara,K. *et al*. (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun*., **4**, 3612.

Yu,Z. *et al*. (2016) CloneCNA: detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data. *BMC Bioinformatics*, **17**, 310.

Zhang,J.D. *et al*. (2017) Detect tissue heterogeneity in gene expression data with BioQC. *BMC Genomics*, **18**, 277.

Zhong,Y. *et al*. (2013) Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, **14**, 89.

Zhong,Y. and Liu,Z. (2012) Gene expression deconvolution in linear space. *Nat. Methods*, **9**, 8–9.

Zinovyev,A. *et al*. (2013) Blind source separation methods for deconvolution of complex signals in cancer biology. *Biochem. Biophys. Res. Commun*., **430**, 1182–1187.

Zuckerman,N.S. *et al*. (2013) A self-directed method for cell-type identification and separation of gene expression microarrays. *PLoS Comput. Biol*., **9**, e1003189.