Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study

Renaud Gaujoux [a], Cathal Seoighe [b,*]

[a] Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, South Africa
[b] School of Mathematics, Statistics and Applied Mathematics, National University of Ireland Galway, Ireland

ARTICLE INFO

Article history:
Available online 10 September 2011

Keywords:
NMF
Microarray
Gene expression
Deconvolution
Sample heterogeneity

ABSTRACT

Heterogeneity in sample composition is an inherent issue in many gene expression studies and, in many cases, should be taken into account in the downstream analysis to enable correct interpretation of the underlying biological processes. Typical examples are infectious diseases or immunology-related studies using blood samples, where, for example, the proportions of lymphocyte sub-populations are expected to vary between cases and controls.

Nonnegative Matrix Factorization (NMF) is an unsupervised learning technique that has been applied successfully in several fields, notably in bioinformatics where its ability to extract meaningful information from high-dimensional data such as gene expression microarrays has been demonstrated. Very recently, it has been applied to biomarker discovery and gene expression deconvolution in heterogeneous tissue samples.

Being essentially unsupervised, standard NMF methods are not guaranteed to find components corresponding to the cell types of interest in the sample, which may jeopardize the correct estimation of cell proportions. We have investigated the use of prior knowledge, in the form of a set of marker genes, to improve gene expression deconvolution with NMF algorithms. We found that this improves the consistency with which both cell type proportions and cell type gene expression signatures are estimated. The proposed method was tested on a microarray dataset consisting of pure cell types mixed in known proportions. Pearson correlation coefficients between true and estimated cell type proportions improved substantially (typically from about 0.5 to approximately 0.8) with the semi-supervised (marker-guided) versions of commonly used NMF algorithms. Furthermore known marker genes associated with each cell type were assigned to the correct cell type more frequently for the guided versions. We conclude that the use of marker genes improves the accuracy of gene expression deconvolution using NMF and suggest modifications to how the marker gene information is used that may lead to further improvements.

1. Introduction

A typical objective in gene expression studies using microarrays or deep sequencing is the identification of genes that are differentially expressed between groups of samples, such as case vs. control or normal vs. tumor tissue. Although sample heterogeneity is widely acknowledged as potentially a substantial confounding factor in such analysis (Cleator et al., 2006; Whitney et al., 2003), it is often discarded, due to the unavailability of data on the composition of the samples. Laboratory techniques such as laser capture micro-dissection, fluorescence-activated cell sorting or flowcytometry exist to separate or quantify constituents from each sample. However, these require extra resources that may not be available in all situations, such as sample quantity, time, technology or funds, beside the fact that manipulating the samples may alter the original gene expression profiles. The capacity to accurately deconvolve gene expression data computationally is, therefore, an attractive alternative. Starting with Venet et al. (2001), many authors proposed different methods and provided insights on how to estimate the cell type/tissue specific signatures or their relative proportions (Zhao and Simon, 2010). Some methods perform partial gene expression deconvolution in the sense that they require that either the signatures (Lu et al., 2003; Wang et al., 2006; Abbas et al., 2009; Clarke et al., 2010), or – estimates of – the mixture proportions (Lähdesmäki et al., 2005; Erkkilä et al., 2010; Shen-Orr et al., 2010) are available. Others perform complete deconvolution where both the cell/tissue signatures and the proportions are estimated directly from the global gene expression data of the heterogeneous samples (Roy et al., 2006; Repsilber et al., 2010).

* Corresponding author. Tel.: +353 91 492343.
E-mail addresses: renaud@cbio.uct.ac.za (R. Gaujoux), cathal.seoighe@nuigalway.ie (C. Seoighe).
URL: http://web.cbio.uct.ac.za/~renaud (R. Gaujoux).

1567-1348/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.08.014

Complete gene expression deconvolution may provide valuable information about the underlying biological processes of interest, particularly in the context of infectious diseases, immunology-related or cancer studies, where whole blood, PBMC or tissue samples are used to compare different phenotypic groups of patients. An accurate estimate of the proportions of constituent cell types in a sample can provide insights into the inflammatory stage of each sample and possibly uncover group-specific patterns, or enable gene expression estimates to be corrected for stromal contamination, a common issue for tumor samples. At the cell-type level, the estimated signatures can be used to extract gene modules, which may reveal cell-type specific gene interactions or pathway activations. Performing complete deconvolution separately on each phenotypic group and looking at the differences in these gene modules between groups may then identify genes or pathways that play a key role in the response to the disease. For this type of analysis to give meaningful results, it is therefore important that the estimated signatures are consistent with the relevant constituting cell-types.

In this paper, we explore the potential benefits of a simple approach that incorporates prior knowledge from marker genes into general algorithms that perform Nonnegative Matrix Factorization (NMF) (Paatero and Tapper, 1994; Lee and Seung, 1999) for the complete gene expression deconvolution problem. NMF is an unsupervised technique that has been successfully applied to a broad range of fields, including bioinformatics, and has proved to be capable of extracting meaningful components from composite data (Devarajan, 2008; Brunet et al., 2004; Pehkonen et al., 2005; Hutchins et al., 2008). Because the NMF theoretical framework is naturally suitable to modeling the problem of complete gene expression deconvolution, many deconvolution methods have been developed within this framework and these often share common algorithmic properties (Venet et al., 2001; Lähdesmäki et al., 2005; Repsilber et al., 2010). One of the drawbacks of standard NMF methods is that the estimation process is completely unsupervised, which does not guarantee that the extracted components are related to the problem of interest. In particular, in the case of gene expression deconvolution, one expects that the cell/tissue signatures exhibit block-like expression patterns, at least for genes that are known to be characteristic of the different cell types. Our proposition is to enforce the recovery of such signature patterns within the estimation process, instead of looking for them a posteriori, with the risk of obtaining biologically inconsistent components. This paper explores the benefit of guiding the estimation in such a way and shows how, on a real dataset, the proposed approach is able to dramatically improve the capacity of standard NMF algorithms to both recover meaningful cell/tissue signatures and accurately estimate their relative mixture proportions.

2. Material and methods

2.1. Data

We evaluated the performance of several NMF methods on the microarray dataset GSE11058 accessible at NCBI GEO database (Barrett et al., 2010). It contains data from a controlled mixture experiment performed by Abbas et al. (2009) to develop their partial deconvolution method. The dataset comprises the gene expression profiles from four pure cell lines of immune origin (Jurkat, IM-9, Raji, THP-1) as well as four different mixtures for which the relative proportions of each cell type are known. Mixtures of cells were performed in triplicate making up a total of 24 arrays (triplicates of each pure cell type and mixture). Because both the pure gene expression profiles and the mixture proportions are available, these data provide a ground truth reference against which the proportions and cell-type expression signatures obtained from complete deconvolution can be assessed. For the latter, the mean expression across pure cell type samples was used as the reference.

We used the normalized gene expression data stored as Series Matrix files available from GEO. This data is normalized by global scaling with Microarray Suite version 5.0 (MAS 5.0) using Affymetrix default analysis settings, with the trimmed mean target intensity of each array arbitrarily set to 500 (e.g. see description page for sample GSM279589). The complete dataset (54675 probesets) was further processed to produce a curated and a full dataset that were subsequently used in the analysis. The curated dataset is limited to the 359 probesets that compose the final basis matrix used in Abbas et al. (2009) to deconvolve white blood cell samples from Systemic Lupus Erythematosus patients, and consists of a set of marker probesets for common immune cell subsets in different states (e.g. resting or activated). The purpose of this dataset is to assess the performance of the deconvolution methods in a setting that is a priori favorable, due the high discriminative power of the probesets (see Supplementary Fig. 1). Moreover, it provides insight about how deconvolution works when only considering a limited number of genes. The full dataset is composed of the 40791 probesets that could be mapped to an Entrez Gene identifier, using the annotation package hgu133plus2.db from bioconductor (Gentleman, 2004), filtering out any built-in Affymetrix control probesets, whose probe IDs start with the prefix `AFFX-`.

2.2. Methods

2.2.1. NMF algorithms

We considered seven NMF algorithms, among which three are guided algorithms that incorporate prior knowledge from marker probesets within the fitting process, using the method described in Section 3. A brief description of each method as well as some details about their implementation follows. Canonical method names are underlined to distinguish them from labeling names.

The method *deconf* was proposed by Repsilber et al. (2010) specifically for performing gene expression deconvolution. It applies an alternating least-square schema to minimize the euclidean distance between the target matrix and the NMF estimate. After each least-square fit, both non-negativity and scaling or sum-to-one constraints are enforced onto the basis and the mixture coefficient matrices.

Algorithm *lee*, which was proposed by Lee and Seung (1999) initially for image recognition, inspired several other NMF algorithms. In this work, we considered the version that minimizes the euclidean distance via iterative multiplicative updates, which are derived from a gradient descent approach. In its original definition, the algorithm only ensures that the non-negativity constraints are satisfied at each iteration. We enforced the sum-to-one constraint on the final fit only, by scaling the columns of the mixture coefficients.

Algorithm *brunet* was developed by Brunet et al. (2004) to perform class discovery in cancer studies, and is an enhancement of Lee's algorithm for minimizing the Kullback-Leibler divergence (Lee and Seung, 2001). One of its particular features is the introduction of a stopping criterion based on the stationarity of the clustering consensus matrix, which makes sense in the context of class discovery, as it indicates that the the model achieved a stationary point for the clusters. However, this criterion is too lax in the case of deconvolution, as it might stop the algorithm too early and prevent further improvements of the estimation accuracy. For this reason, we instead used a stopping criterion based on the stationarity of the objective function, which indicates that the approximation of the target matrix would not improve significantly with further iterations. As per method *lee*, scaling of the final mixture coefficient matrix ensures the result satisfies the sum-to-one constraint on the proportions.

The method *nsNMF* was designed (Pascual-Montano et al., 2006) for performing bi-clustering of microarray data, and introduces a constant smoothing matrix into the model in order to obtain sparser results. For this method too, we used the stopping criterion based on the stationarity of the objective function instead of the clustering consensus matrix. The mixture proportions are obtained as the product of the smoothing matrix by the final mixture coefficient matrix, with a final column scaling to satisfy the sum-to-one constraint.

Methods *G-lee*, *G-brunet* and *G-nsNMF* are modifications we made from the methods *lee*, *brunet* and *nsNMF*, respectively, that take into account prior knowledge of markers for each cell type using the strategy described in Section 3. In the remainder of the paper, we refer to these last three methods as the *guided* methods, and sometimes substitute the prefix "G" for a numerical suffix that specifies the number of markers used in the fitting process (e.g. *lee-5* refers to the method *G-lee* that uses 5 markers per cell type). Different NMF models were estimated for each of these methods using an increasing number of markers, precisely {1–10, 15, 20, 25} and {5, 10, 20, 30, 40, 50, 70, 90, 120, 200} for the curated and the full datasets, respectively.

An important point to bear in mind is that the first four methods actually do require and use the same prior knowledge as the guided methods, that is a set of marker genes for each cell-type. Indeed such methods return estimated basis components in an unpredictable order and unlabeled. Marker genes are used to map each basis component to one of the real cell type signatures (see Section 2.2.3). Hence, strictly speaking, all methods are supervised, some at a final mapping stage, others during the estimation step.

2.2.2. Marker selection

We selected a set of marker probesets for each cell type based on the differences in gene expression observed between the pure samples. For each probeset, we computed a standard *t*-test statistic between the samples from the cell type in which the gene was most highly expressed and the second most highly expressing cell type (Abbas et al., 2009), and defined as markers the probesets with a *p*-value less than 0.05 and a $\log_2$ fold change greater than 1.5. The *p*-values were computed on the $\log_2$ transformed data, using a two-sided *t*-test with equal variance. Thus, markers were selected for which the expression level in the most highly expressing cell type was significantly greater than the expression level in the next most highly expression cell type. Table 1 shows the total number of markers per cell type obtained for each dataset. In the case of the full dataset, only the top 300 markers of each cell type were used in the subsequent analysis. This is to limit the number of false positives (cf. the *q*-values in Table 1), in addition to the fact that, in practice, it is unrealistic to require a very large number of markers for each cell type of interest.

These markers were used by the guided methods to enforce the expected expression block pattern on the estimated signatures (cf. Section 3), and by the non-guided methods to a posteriori map the estimated signatures to the real cell types (cf. Section 2.2.3).

Table 1
Total number of markers per cell type for each dataset. For the full dataset, the *q*-values estimate for each cell type the proportion of false positives expected in the top 300 markers.

Datasets	Cell types			
	Jurkat	IM-9	Raji	THP-1
Full	733	562	437	1294
q-Value (300)	0.028	0.058	0.062	0.004
Curated	21	17	20	26

2.2.3. Cell type mapping

As already stated, all methods require and use a set of markers at some stage of the deconvolution process. The guided methods *G-brunet*, *G-lee* and *G-nsNMF* use the markers to enforce cell type specific expression patterns on each basis component. This means that each component is de facto associated with a given cell type and no final mapping stage is necessary. For the non-guided methods *deconf*, *brunet*, *lee* and *nsNMF* however, the order of the components is not known a priori and these need to be mapped heuristically and a posteriori to one of the cell types. Hence the mapping process is critical in this case as it provides all their meaningfulness to the results: trying to estimate proportions is meaningless if these cannot be reliably associated with the correct cell types.

Repsilber et al. (2010) applied a majority count decision rule to assign the estimated components from two cell types. We extended the principle to make it work robustly for any number of cell types. The mapping strategy consists in iteratively associating each component with the cell type with the maximum percentage of consistent markers. More precisely, we first build a predicted map that assigns each marker to the estimated component that expresses it the most and compute the contingency table of this map with the theoretical map built from the marker list. Each entry in the contingency table is the number of markers that are consistent between a given component and a given cell type. The columns of the contingency table are scaled to sum to one in order to obtain the percentages of markers from each cell type that are consistent with each component. The component and the cell type that achieve the maximum percentage of consistent markers are mapped together and removed from the contingency table. The mapping is repeated until all components have been assigned to a cell type. The components obtained from the non-guided methods were assigned using this strategy with the complete set of markers, as this is expected to give more robust mappings.

2.2.4. Implementation details

All computations were done within R (R Development Core Team, 2011), using the package NMF (Gaujoux and Seoighe, 2010), which provides a general framework for running, developing and testing NMF algorithms. We used the built-in optimized version of the methods *brunet*, *lee* and *nsNMF*, only changing the stopping criterion as described in Section 2.2.1. A maximum of 2000 iterations was allowed. The guided methods *G-brunet*, *G-lee* and *G-nsNMF* were implemented upon their respective non-guided versions, by enforcing inclusion of the marker patterns on the basis matrix after each iteration. The method *deconf* was implemented within the same framework by wrapping the function provided the R package `deconf` available in Supplementary data of Repsilber et al. (2010).

All methods need to be initialized with a starting point, i.e. an initial NMF model. This is randomly chosen by drawing the entries of the basis and mixture coefficient matrices from a uniform distribution $\mathcal{U}[0, \max(X)]$, where X is the global gene expression matrix. The mixture coefficients are then scaled to satisfy the sum-to-one constraint. Given that none of the methods have established global convergence properties, all NMF estimates were obtained as the fit that achieved the least residual error from 200 runs, each one using a different random initialization. To avoid biasing the comparisons by the choice of different starting points, and out of concern for reproducible research (Hothorn and Leisch, 2011), we fixed the random seed to a common value before each set of runs (seed = 123456). This allows the package NMF to guarantee that each set of runs uses a common sequence of random initializations, generated by independent random streams (L'Ecuyer et al., 2002; L'Ecuyer and Leydold, 2005). See Appendix A for detailed information on the R installation used to generate the results.

3. Theory

Although the relationship between the expression levels of pure and mixed samples is known not to be strictly linear, previous work on gene expression deconvolution showed that the linearity assumption is reasonnable (Shen-Orr et al., 2010). Hence, the complete gene expression deconvolution problem is commonly formulated as an extended linear model. In this paper, we use the following Nonnegative Matrix Factorization (NMF) theoretical framework.

Given a nonnegative $n \times p$ matrix X, Nonnegative Matrix Factorization aims at finding an approximation

$$X \approx WH, \tag{1}$$

where W, H are $n \times r$ and $r \times p$ non-negative matrices, respectively, and the factorization rank r is often such that $r \ll \min(n,p)$.

In essence, Eq. (1) simply states that each column of X (i.e. the observed features of each sample) is approximated by a non-negative linear combination of the columns of W (i.e. the basis components), where the coefficients are given by the corresponding column of H (i.e. the mixture coefficients). If one imposes moreover that the columns of H sum to one, an NMF model such as (1) may be directly interpreted in terms of gene expression deconvolution: the matrix X represents the global gene expression matrix from heterogeneous samples (e.g. blood or PBMCs), the columns of the matrix W correspond to specific gene expression signatures of the cell types (e.g. T-cells, Monocytes), and each column of the matrix H provides the proportions of each cell type in the corresponding sample.

Classical NMF algorithms use iterative optimization methods to minimize an objective function that measures the distance between the target global gene expression matrix and its NMF estimate. Common objective functions are based on the Frobenius norm or the Kullback-Leibler divergence (Lee and Seung, 2001; Cichocki et al., 2008). Variations on the Eq. (1) or the optimization problem exist in order to take into account some a priori knowledge about the data or the solution (Hoyer, 2004; Pascual-Montano et al., 2006). One example of such variations is the sum-to-one constraint we imposed on the mixture coefficient matrix H in order to represent relative proportions, instead of absolute counts.

Our proposition is to impose another set of constraints on the signature matrix W, with the objective of estimating more stable and meaningful cell type signatures. This should, in turn, improve the estimation of the mixture proportions. In order to achieve this we use a set of marker genes, each one of which is known to be – almost – exclusively expressed by just one of the cell types. We therefore want to constrain the rows of the signature matrix W that correspond to each marker gene so that all entries are zero except one. In this way we associate a priori each gene expression signature (i.e. each column of W) to a given cell type.

Many NMF algorithms such as *lee*, *brunet*, and *nsNMF*, implement gradient-descent methods using iterative multiplicative updates. These define the next iterate value of each factor (W and H) as its element-wise product by another matrix, chosen to ensure – at least – that the objective function is non-increasing (Berry et al., 2007). Therefore, theoretically, for this kind of algorithm, enforcing block patterns on the initial signatures guarantees their persistence alonog all iterations. In practice however, due to adjustments commonly made to avoid numerical difficulties, one may require to enforce the block patterns after each iteration. Hence at initialization and after each iteration of the chosen NMF algorithm, each cell type signature has the values corresponding to markers of other cell types set to zero. The values for its own markers are left free to be updated by the algorithm's own iterative schema.

On the other hand, the method *deconf* is based on an alternating least-squares strategy. Rigourously constraining block expression patterns for this kind of algorithm, requires more sophisticated approaches such as projected-gradient methods (Lin, 2007). In fact, the method *deconf* already imposes the non-negativity and sum-to-one constraints using a heuristic, whose implications on the objective value are not clear; all the more so if marker constraints are added. Therefore, for the purpose of this paper, we incorporated such constraints only into the multiplicative NMF algorithms considered, viz. *lee*, *brunet* and *nsNMF*. However, the performances achieved by *deconf* on the full dataset suggests that adapting this algorithm to make use of markers could potentially be fruitful.

4. Results and discussion

All analyses were performed on both the curated and the full datasets. Since we are interested in complete deconvolution, the pure samples were excluded from the expression matrix, which mimics the realistic situation in which expression data for the pure samples are not available. Fig. 1 shows the mean absolute differences (mAD) between the true and estimated proportions achieved by each method for a varying number of markers, on both datasets. The values achieved by the guided methods are plotted with bullets, those achieved by the non-guided methods with single solid diamonds at abscissa 0. Note that this abscissa choice is for plotting purposes only, and does not reflect the actual usage of markers by these methods (cf. Section 2.2.1), since all of these methods do make use of markers to identify components with cell types. The benefit of using markers to guide the fitting process is clear on both datasets. Indeed, the guided methods achieve significantly lower mAD values than their respective non-guided versions, and this for any number markers, meaning that enforcing marker patterns on the basis components improves the accuracy of the mixing proportions estimates. The improvement in accuracy is particularly striking in the case of the curated dataset, where using a single marker per cell type already improves dramatically the estimation of the mixture proportions, specially for the methods *brunet* and *lee*. The best accuracy was achieved by the method *brunet-7* (mAD = 0.05). In Figs. 2 and 3(a–b) we highlight the differences between the estimates obtained from the methods *brunet*, *brunet-7* and *deconf* on the curated dataset. For completeness, we show in Supplementary Figs. 2–4 the plots obtained for each method and each number of markers. These are animated plots which highlight the effect of increasing the number of enforced markers.

Fig. 2(a and c) shows scatter plots of the estimated versus the true mixture proportions for *brunet* and *brunet-7*, respectively. On each plot, the colors distinguish between the four cell types, the global pearson correlation coefficient r is indicated at the bottom right, and the values in parenthesis within the legend indicate the pearson correlation coefficients computed for each cell type separately. The proportion estimates from the guided method *brunet-7* are much more accurate (mAD = 0.05) and highly correlated with the true proportions (r = 0.91), than the estimates obtained by its non-guided version *brunet* (mAD = 0.208 and r = 0.52).

The heatmaps in Fig. 2(b and d) supports the same conclusions. These show the expression of all the marker probesets across the estimated signatures. The assigned cell type is indicated at the bottom of each signature. To emphasize the differences between cell types, the expression levels were scaled in each row separately into relative percentages of expression. The color palette ranges from light yellow to dark red for, respectively, 0% and 100% of expression. The markers are ordered by increasing p-value within their respective reference cell type. The colored annotation columns on the left hand side of the heatmap indicate which estimated cell type expresses each marker the most highly. Markers are colored according to their respective true cell type, using the same color

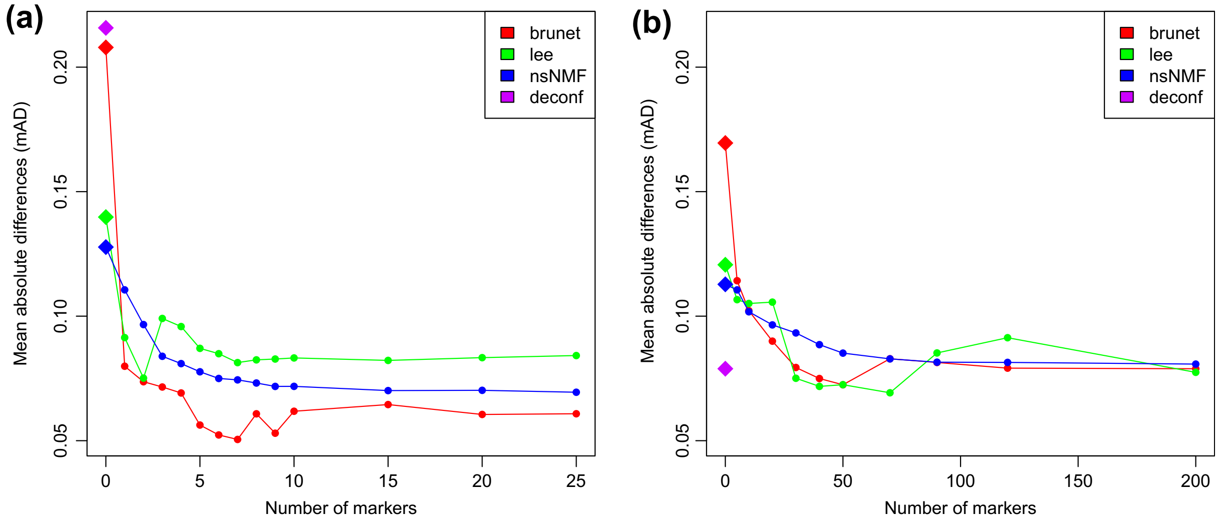

Fig. 1. Mean absolute differences between the true and estimated proportions achieved on the curated dataset (a) and the full dataset (b). Solid diamonds and bullets indicate the non-guided and guided methods, respectively.

code as the scatter plots. Hence, a method correctly recovers the markers of a given cell type if the associated annotation column consists of a single monochromatic block. As a result of the strategy used to guide the algorithm, the enforced markers of each cell type show a 100%-value in their corresponding components, and 0 elsewhere.

Fig. 2(d) shows that using markers successfully resolved the inconsistent signatures obtained for IM-9 and THP-1 by the standard method *brunet* (Fig. 2(b)). In fact, independently of the number of markers being enforced, all guided methods estimated signatures which exhibit the expected block pattern (Supplementary Figs. 2–4). The recovered signatures are not only more defined than the one estimated by the standard methods, but they are also biologically more meaningful as most of the markers are highly expressed by the correct cell type. Despite the fact that a single marker per cell-type was sufficient to guide the algorithm towards relevant cell-type signatures, more markers were needed to gradually improve the accuracy; up to a certain point after which the proportion estimates seem to become biased, although being more precise. For example, the proportion of IM-9 seems to be systematically under-estimated by *brunet-25*, resulting in a compensatory over-estimation of the other proportions, that however appears evenly divided amongst the different cell-types (Supplementary Fig. 2).

Fig. 3(b) indicates that a similar inconsistency issue affects the signature estimated by the method *deconf* (mAD = 0.216).

In this case, the estimated signature assigned to THP-1 expresses at a high level most of the markers for Jurkat as well; somehow even more clearly than the markers for THP-1 itself. We noticed that swapping the signatures assigned to these two cell-types improved the accuracy (mAD = 0.172), which suggests that they were mis-assigned (Supplementary Fig. 5). However, this does not change the signatures themselves, which remain inconsistent with the real underlying cell types, limiting the method's ability to properly estimate the mixture proportions (Fig. 3(a)).

As far as the method *nsNMF* is concerned, the mAD plots in Fig. 1 show that, when non guided, it achieved consistently lower mAD values than *brunet* and *lee*, but benefited relatively moderately from the usage of the markers compared to the two latter methods. Probably due to its extra sparsity constraint however, its guided version estimates some proportions very accurately (IM-9 and Raji), but others with a systematic bias (Supplementary Fig. 4). On the other hand, the method *lee* seems to be somehow more sensitive to the variation in the number of markers, compared to *brunet* and *nsNMF*. This could be explained by fundamental differences in the objective functions these methods optimize. Indeed, *brunet* and *nsNMF* minimize the Kullback-Leibler divergence, which is based on log-differences, whereas *lee* minimizes the euclidean distance, which is based on square-differences, making it more sensitive to deviations, in particular to those that arise from the enforcement of the marker expression patterns.

Fig. 1 shows that the method *deconf* achieved a remarkable accuracy when applied to the full dataset (mAD = 0.079), and is only outperformed by the guided methods *G-lee* and *G-brunet* when using an appropriate number of markers. Fig. 3(c) indicates that the global pearson correlation is high ($r = 0.82$) and that all cell types except THP-1 were recovered with a correlation greater than 0.9. Along the same lines, the heatmap in Fig. 3(d) reveals that the estimated signature for THP-1 is particularly inconsistent with its associated markers, all of them being mostly expressed by the component assigned to Jurkat. Because the other three cell types are relatively well recovered, the estimation of the mixture proportions would not be completely hampered, specially with the presence of the sum-to-one constraint. On the other hand, Fig. 4d shows that, when guided by 30 markers per cell type, the method *lee* recovers cell type signatures that are extremely consistent with the real cell types, while estimating the mixture proportions with a similar accuracy (mAD = 0.075). In particular, enforcing these markers improves the correlation of the estimated proportions of THP-1 from 0.56 to 0.79 (cf. Fig. 4(a and c)). The plots obtained for each method and each number of markers are all shown in Supplementary Figs. 6–8.

On a more general level, these results raise the main potential caveat of using the non-guided methods. Their performance heavily relies on their ability to recover meaningful cell signatures without being supervised. Given the noise inherent in gene expression data and other possible confounding factors, this is not guaranteed to succeed, especially when estimating more than two cell types. Moreover, the interpretation of their results also depends on the mapping heuristic that assigns the estimated components to a real cell type. As an example, Table 2 shows the percentages of consistent markers obtained by the method *deconf* on the full dataset. All the markers were used from each cell type. The columns correspond to the real cell types, the rows to the estimated signatures,

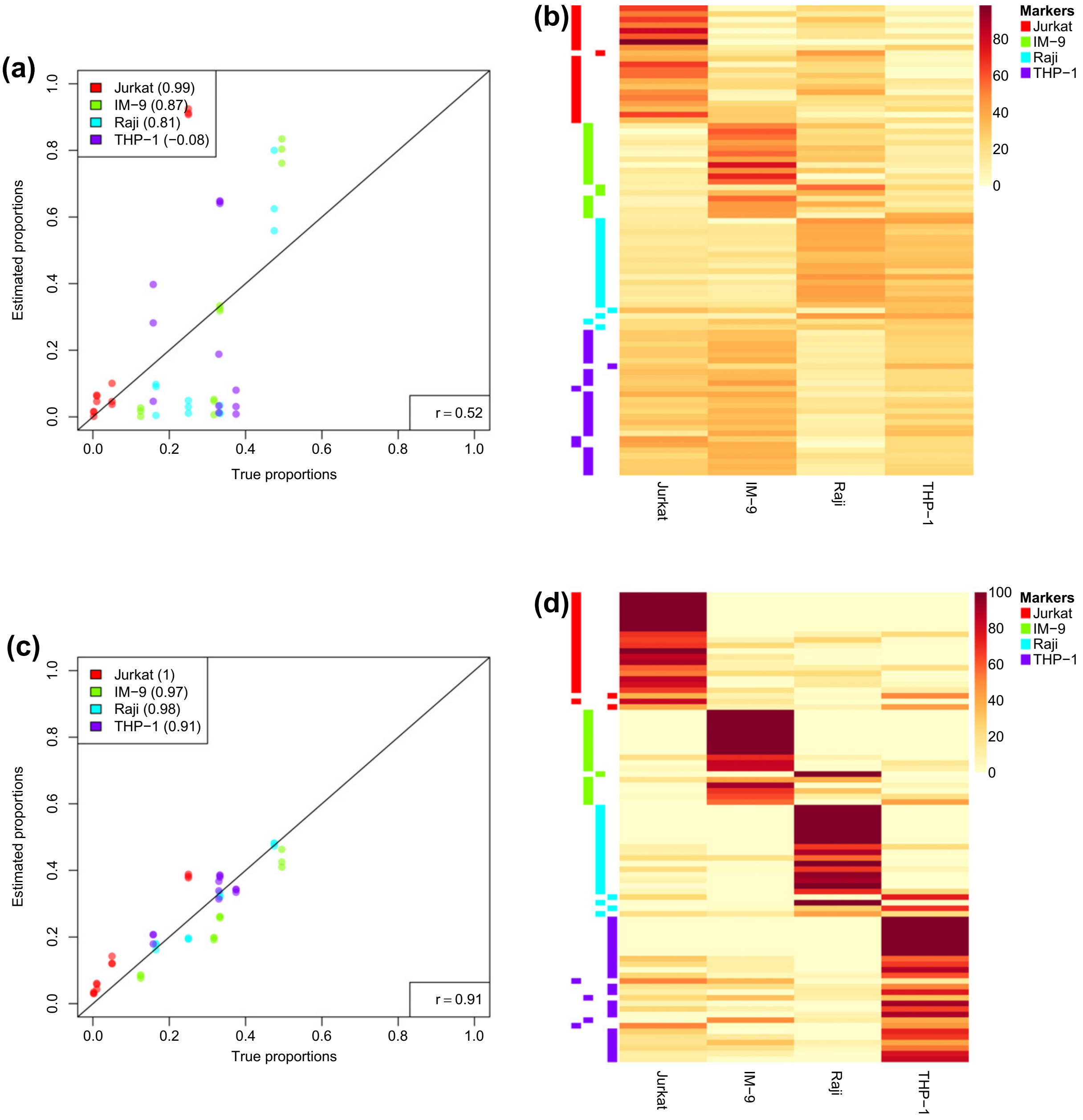

Fig. 2. Estimated proportions vs. true proportions and heatmaps of the estimated cell type signatures on the curated dataset by the methods *brunet* (a–b) and *brunet-7* (c–d). The heatmaps show the expression of all of the marker probesets across the estimated signatures. The assigned cell type is indicated at the bottom of each signature. Rows were scaled separately into relative percentages of expression. The markers are ordered by increasing *p*-value within their respective reference cell type. The colored annotation columns on the left hand side of the heatmap indicate which estimated cell type expresses each marker the most highly.

and the table reads e.g. 88.33% of the Jurkat markers are most expressed in signature number 2. The iterative mapping strategy described in Section 2.2.3 assigns the signatures in the following order: 2 → Jurkat, 1 → Raji, 3 → IM-9 and finally 4 → THP-1, despite the fact that only 4.33% of the markers for THP-1 are most highly expressed on signature 4. One could argue in this case that the cell type THP-1 was not recovered at all, and rather that it was distributed between signatures 2 and 3. In this context, the strategy that consists of enforcing marker expression patterns on the signatures presents two main advantages. First it does not require choosing a mapping heuristic since each signature is de facto assigned to a cell type up front. Second, it has the potential to guide the algorithm toward meaningful and consistent signatures, when the appropriate number of markers are used.

However, we acknowledge that the approach as presented in this paper has some limitations. In order to maximize the cell type specific signals carried by the marker probesets, we selected these based on pure samples from the same data. Although our purpose was to explore the potential of the guiding approach, this would not be possible in a real setting where these samples would not be available. The analysis would gain in being reproduced using markers that are defined independently from the data, such as the set of markers characterized by Abbas et al. (2005) for different immune subsets. Moreover, assessing the performance of this semi-supervised approach on more complex real gene expression datasets, from whole blood or PBMC samples, would provide better insights on its applicability to clinical research.

Besides the choice of the discriminative markers themselves, the semi-supervised approach requires a choice of how many markers to enforce. Using too few markers might limit the method's ability to separate all of the underlying cell-types. It also appears that enforcing too many markers is detrimental to the

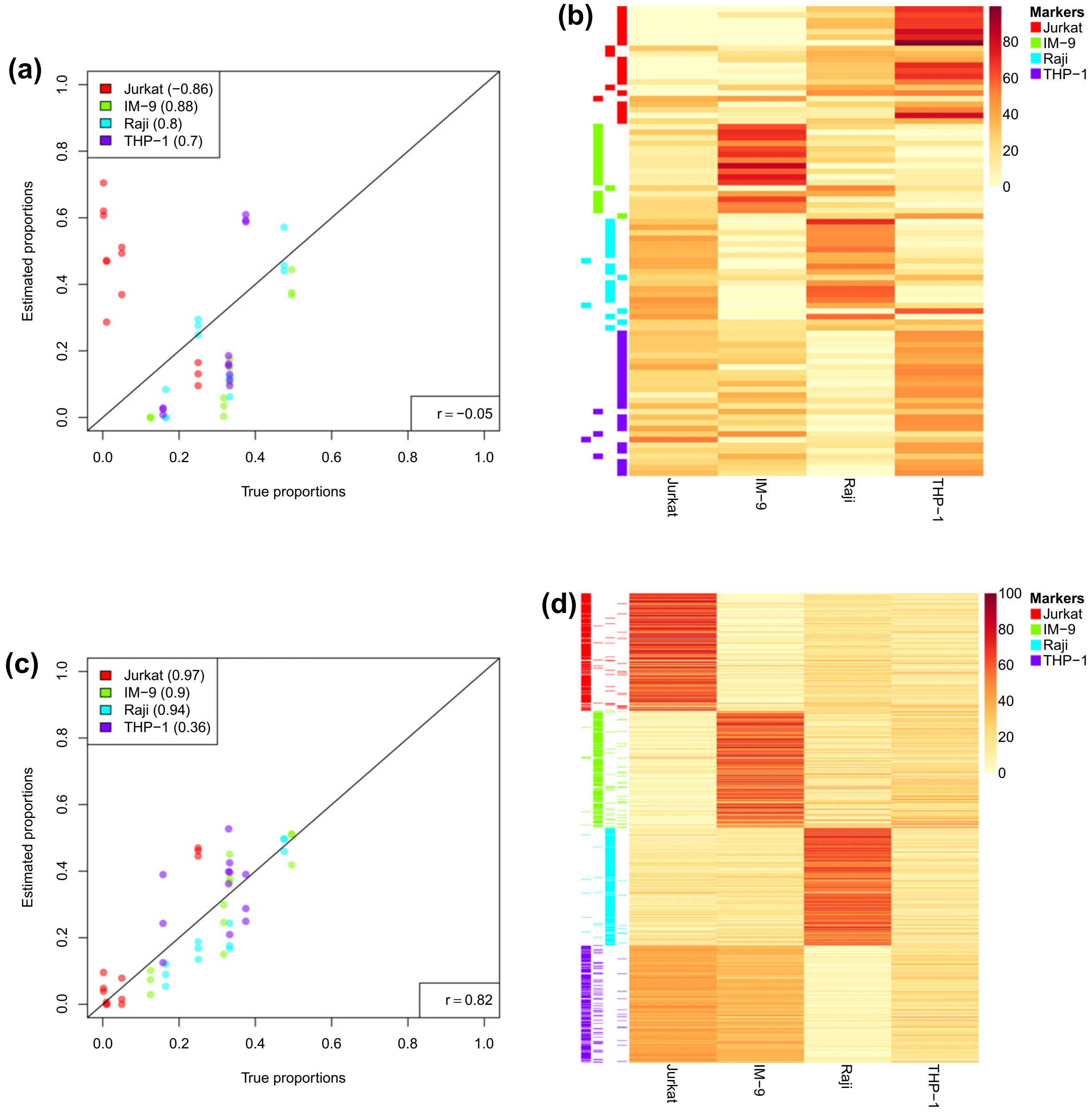

Fig. 3. Estimated proportions vs. true proportions and heatmaps of the estimated cell type signatures by the method *deconf* on the curated dataset (a–b) and the full dataset (c–d). See Fig. 2 for a more detailed description of the plots.

global accuracy (Fig. 1(a)). As part of a preliminary analysis, we had fitted NMF models on the complete data, i.e. including the 12 pure samples in the full target matrix. The idea was to investigate how the different methods make use of the actual cell type signatures, when these are already part of the data. We observed that the method *brunet* particularly, and to a lesser extend the methods *lee* and *nsNMF*, estimated the mixture proportions with high accuracy, while all of them perfectly recovered every single cell type markers (Supplementary Figs. 10–12 and 14–16). In comparison, the method *deconf* performed worse, estimating messier signatures and less accurate proportions (Supplementary Figs. 13 and 17). This made the three former algorithms good candidates for implementing our guided deconvolution approach. However, in the presence of the pure samples, using markers to enforce block patterns was detrimental (Supplementary Fig. 5). We explained the observed degradation in accuracy, by the fact that poorly fitting the pure sample profiles would result in large residuals, specially due to their sparse structure. Non-guided methods, having no constraint imposed on the signature profiles, can finely optimize them, which leads to accurate estimates of the mixture coefficients. The block patterns enforced by the guided methods are then counterproductive in this case, as they constitute a rough approximation compared to the fine grained information provided by the true pure samples. Rather than suggesting that the use of markers could not be beneficial at all, these results confirmed the fact that incorporating prior knowledge of the expression profiles of cell types that make up the samples is an effective way to guide NMF based gene expression deconvolution algorithms towards meaningful solutions.

Possible improvements to the approach include refining the strategy used to enforce the markers in this study, in order to make it more robust with regards to both the choice and the number of

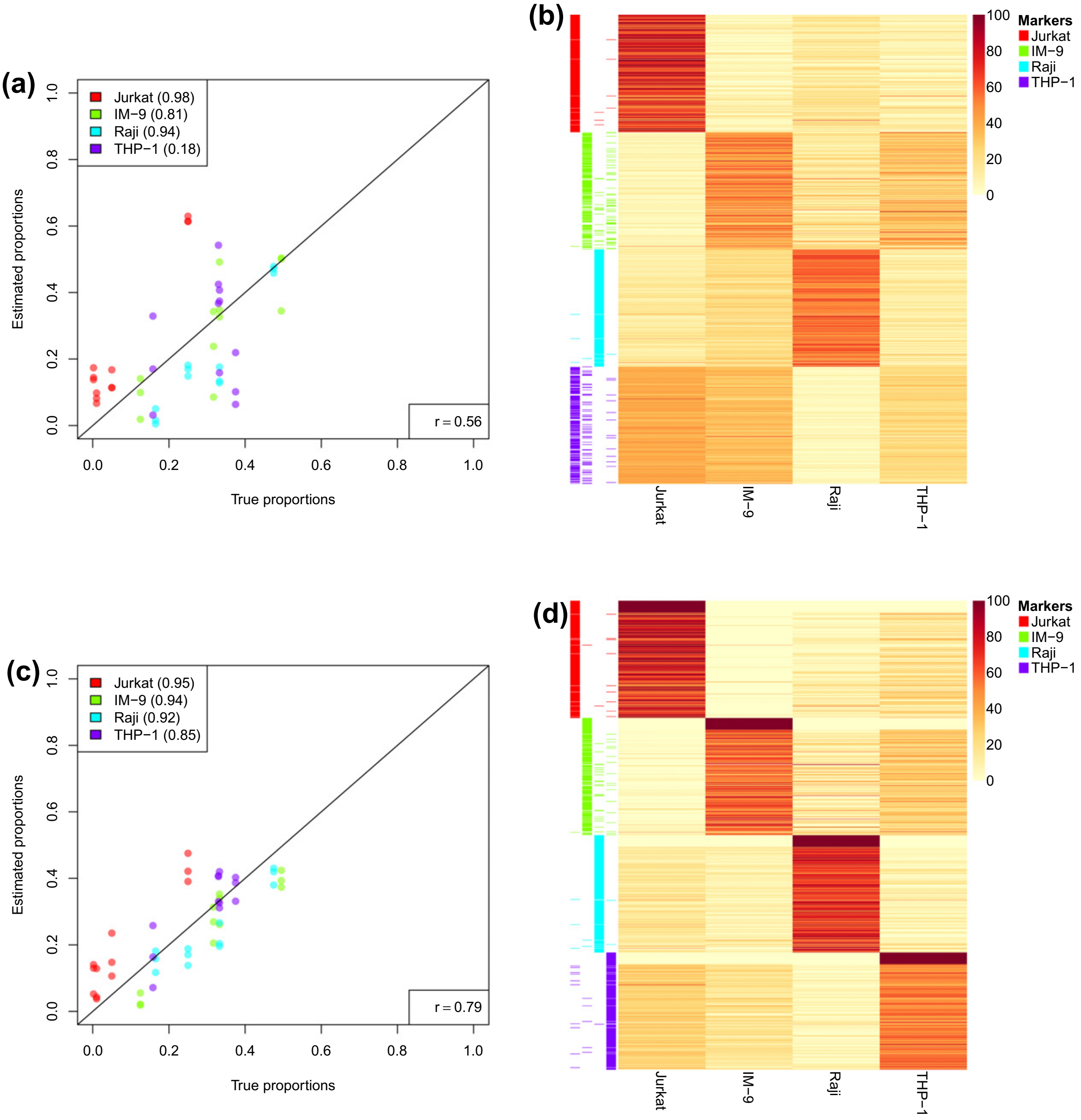

Fig. 4. Estimated proportions vs. true proportions and heatmaps of the estimated cell type signatures on the full dataset by the methods *lee* (a–b) and *lee-30* (c–d). See Fig. 2 for a more detailed description of the plots.

Table 2
Percentages of consistent markers for the method *deconf* applied on the full dataset.

	Jurkat	IM-9	Raji	THP-1
1	4.67	6.67	94.00	0.00
2	88.33	2.00	1.67	77.67
3	2.00	82.00	2.00	18.00
4	5.00	9.33	2.33	4.33

markers. For instance, the constraints currently impose zero expression of marker genes in the signatures of cell types with which they are not associated. This might not be the case in practice for many true cell-type specific markers, although they would a priori be good guiding probeset candidates. Hence, the constraints could be slightly relaxed, by allowing the cell-type signatures to express, to some extent, the markers from another cell-type. This should result in better fitting to the data – even in presence of pure samples, while maintaining the desired marker block patterns. The whole estimation might also be improved by incorporating within the fitting process the sum-to-one constraints on the mixture proportions. This constraint is important because it introduces negative correlations between the proportion estimates, which reduces the search space to mixture coefficients that are physically more meaningful in terms of relative proportions. Finally, techniques to assess the quality of the estimated proportions and cell-type signatures in a real setting should also be investigated. In particular, these could be used to test the discriminative power of the markers and implement a data driven procedure to extract an optimal set of markers for deconvolution.

5. Conclusion

Complete gene expression deconvolution is an attractive alternative to expensive laboratory techniques, to both estimate cell/

tissue proportions and disentangle their specific signals in heterogenous samples, leading to improved interpretations. Performing deconvolution on an entire gene expression dataset is desirable as it may reveal key gene modules or pathways involved in the biological process of interest. However, extracting meaningful cell-type signatures from such high-dimensional data is a difficult problem. In this paper we explored how complete gene expression deconvolution using Nonnegative Matrix Factorization algorithms could be enhanced by enforcing known markers to follow an expected block pattern in the estimated cell type signatures. This approach of incorporating such prior knowledge within the fitting process is promising, as using a small number of markers already greatly improved the accuracy of the mixture proportion estimates. Furthermore, when an appropriate number of markers were used, guided algorithms recovered more meaningful cell type expression signatures. Future work will consist in enhancing the usage of marker information as well as developing tools to help implement and assess gene expression deconvolution methods.

Appendix A. R session information

- R version 2.12.1 (2010-12-16), `x86_64-pc-linux-gnu`.
- Locale: `LC_CTYPE=en_ZA.utf8`, `LC_NUMERIC=C`, `LC_TIME=en_ZA.utf8`, `LC_COLLATE=en_ZA.utf8`, `LC_MONETARY=C`, `LC_MESSAGES=en_ZA.utf8`, `LC_PAPER=en_ZA.utf8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_ZA.utf8`, `LC_IDENTIFICATION=C`.
- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, utils.
- Other packages: AnnotationDbi 1.12.0, bigmemory 4.2.11, Biobase 2.10.0, cacheSweave 0.4-5, codetools 0.2-8, colorspace 1.1-0, DBI 0.2-5, deconf 1.0, digest 0.5.0, doMC 1.2.2, filehash 2.1-1, foreach 1.3.2, hgu133plus2.db 2.4.5, iterators 1.0.5, multicore 0.1-5, NMF 0.5.99, org.Hs.eg.db 2.4.6, RColorBrewer 1.0-5, Rcpp 0.9.6, RSQLite 0.9-4, rstream 1.3.1, stashR 0.3-3, synchronicity 1.0.9, xTable 1.5-6.
- Loaded via a namespace (and not attached): tools 2.12.1.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2011.08.014.

References

Abbas, A.R., Baldwin, D., Ma, Y., Ouyang, W., Gurney, A., Martin, F., Fong, S., van Lookeren Campagne, M., Godowski, P., Williams, P.M., Chan, a.C., Clark, H.F., 2005. Genes and Immunity 6, 319–331.

Abbas, A.R., Wolslegel, K., Seshasayee, D., Modrusan, Z., Clark, H.F., 2009. PloS One 4, e6098.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.a., Phillippy, K.H., Sherman, P.M., Muertter, R.N., Holko, M., Ayanbule, O., Yefanov, A., Soboleva, A., 2010. Nucleic Acids Research 39, 1005–1010.

Berry, M., Browne, M., Langville, A.N., Pauca, P., Plemmon, R., 2007. Computational Statistics and Data Analysis.

Brunet, J.-P., Tamayo, P., Golub, T.R., Mesirov, J.P., 2004. Proceedings of the National Academy of Sciences of the United States of America 101, 4164–4169.

Cichocki, A., Zdunek, R., Amari, S.-i., 2008. IEEE Signal Processing Magazine 25, 142–145.

Clarke, J., Seo, P., Clarke, B., 2010. Bioinformatics (Oxford, England) 26, 1043–1049.

Cleator, S.J., Powles, T.J., Dexter, T., Fulford, L., Mackay, A., Smith, I.E., Valgeirsson, H., Ashworth, A., Dowsett, M., 2006. Breast Cancer Research: BCR 8, R32.

Devarajan, K., 2008. PLoS Computational Biology 4, e1000029.

Erkkilä, T., Lehmusvaara, S., Ruusuvuori, P., Visakorpi, T., Shmulevich, I., Lähdesmäki, H., 2010. Bioinformatics 26, 2571–2577.

Gaujoux, R., Seoighe, C., 2010. BMC Bioinformatics 11, 367.

Gentleman, R.C., 2004. Genome Biology 5.

Hothorn, T., Leisch, F., 2011. Briefings in Bioinformatics.

Hoyer, P., 2004. The Journal of Machine Learning Research 5, 1457–1469.

Hutchins, L.N., Murphy, S.M., Singh, P., Graber, J.H., 2008. Bioinformatics (Oxford, England) 24, 2684–2690.

Lähdesmäki, H., Shmulevich, L., Dunmire, V., Yli-Harja, O., Zhang, W., 2005. BMC Bioinformatics 6, 54.

L'Ecuyer, P., Leydold, J., 2005. R News 5, 16–20.

L'Ecuyer, P., Simard, R., Chen, E., Kelton, W., 2002. Operations Research 50, 1073–1075.

Lee, D.D., Seung, H.S., 1999. Nature 401, 788–791.

Lee, D., Seung, H., 2001. Advances in Neural Information Processing Systems.

Lin C.-j., 2007. Projected Gradient methods for Non-negative Matrix Factorization. Technical Report.

Lu, P., Nakorchevskiy, A., Marcotte, E.M., 2003. Proceedings of the National Academy of Sciences of the United States of America 100, 10370–10375.

Paatero, P., Tapper, U., 1994. Environmetrics 5, 111–126.

Pascual-Montano, A., Carazo, J.M., Kochi, K., Lehmann, D., Pascual-marqui, R.D., 2006. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 403–415.

Pehkonen, P., Wong, G., Törönen, P., 2005. BMC Bioinformatics 6, 162.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Repsilber, D., Kern, S., Telaar, A., Walzl, G., Black, G.F., Selbig, J., Parida, S.K., Kaufmann, S.H.E., Jacobsen, M., 2010. BMC Bioinformatics 11, 27.

Roy, S., Lane, T., Allen, C., Aragon, A.D., Werner-Washburne, M., 2006. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology 13, 1749–1774.

Shen-Orr, S.S., Tibshirani, R., Khatri, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T., Sarwal, M.M., Davis, M.M., Butte, A.J., 2010. Nature Methods 7, 287–289.

Venet, D., Pecasse, F., Maenhaut, C., Bersini, H., 2001. Bioinformatics 17, S279.

Wang, M., Master, S.R., Chodosh, L.a., 2006. BMC Bioinformatics 7, 328.

Whitney, A., Diehn, M., Popper, S., Alizadeh, A., Boldrick, J., Relman, D., Brown, P., 2003. Proceedings of the National Academy of Sciences of the United States of America 100, 1896.

Zhao, Y., Simon, R., 2010. Genome Medicine 2, 93.