An introduction to gene expression deconvolution and the *CellMix* package

A Comprehensive Framework for Gene Expression Deconvolution

Renaud Gaujoux

April 2, 2013

Abstract

This vignette motivates and describes the functionalities of the *CellMix* package, an R package for performing gene expression deconvolution analysis. The package defines a general framework to apply, develop and test gene expression deconvolution methods. It incorporates, generalises and extends the set of tools we implemented when developing a semi-supervised approach to this problem, and includes several other previously published algorithms, hence facilitating their application and comparison. A special focus is drawn on lists of cell/tissue marker genes, which are very valuable resources as they may be used not only as prior knowledge or *post-hoc* independent validation data by deconvolution algorithms, but also as gene sets in classical enrichment analysis. We envisage that this package will provide the bioinformatics research community with an easy to use and flexible platform for working with deconvolution methods, and cell heterogeneity in omics data in general.

This vignette aims at providing a general background on gene expression deconvolution, motivating the *CellMix* package, as well as describing its main features. It serves as supplementary material for the following article:

Renaud Gaujoux et al. "CellMix: A Comprehensive Framework for Gene Expression Deconvolution". In: *submitted* (2012)

Documentation, practical examples, and sample analyses can be found online at:
http://web.cbio.uct.ac.za/~renaud/CRAN/web/CellMix

Contents

1 Introduction and Objectives

Our work on semi-supervised gene expression deconvolution (Gaujoux et al. 2011) revealed several challenges that researchers must face when developing or simply applying gene expression deconvolution methods. These challenges are essentially related to the availability and usability of benchmark expression datasets, cell/tissue specific quantitative signatures, lists of marker genes and/or deconvolution algorithms. Most of the time, one or more of these items will be required, whether the objective is to deconvolve a specific global expression dataset or design new methodologies.

1.1 Computational or *in-silico* gene expression deconvolution

Starting with Venet et al. (2001), many authors provided insights into how to estimate the cell type/tissue specific signatures and/or relative cell type proportions from global gene expression measurements, e.g., by microarray or RNA-seq, and proposed a variety of methods to do so (Zhao et al. 2010). These methods are of two different types, distinct with respect to the input data they require:

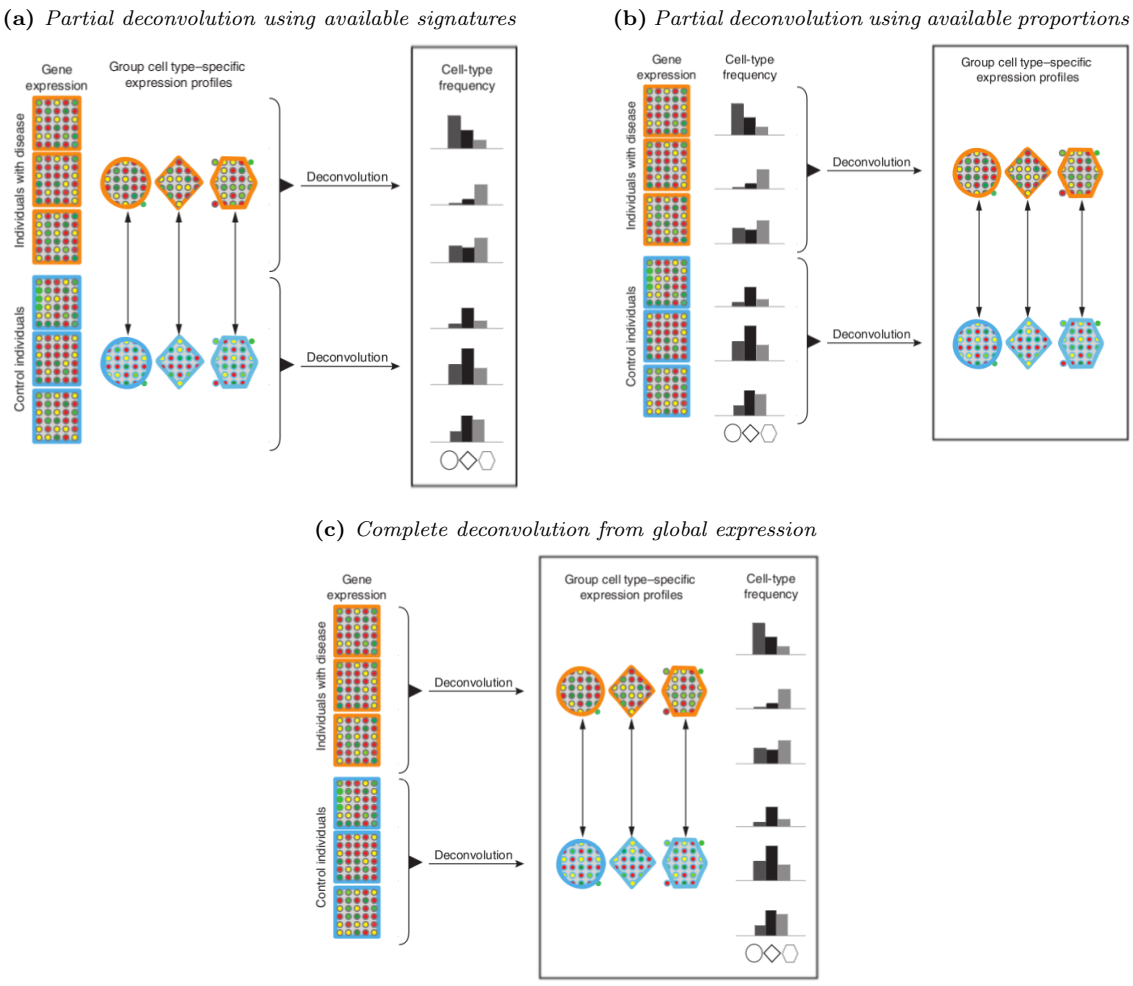
- the *partial* gene expression deconvolution methods require that either cell type-specific signatures (Abbas et al. 2009; Clarke et al. 2010; Gong et al. 2011; Lu et al. 2003; Wang et al. 2006), or mixture proportions (Erkkila et al. 2010; Lähdesmäki et al. 2005; Shen-Orr et al. 2010; Stuart et al. 2004) are available;
- the *complete* deconvolution methods estimate both the cell/tissue signatures and the proportions directly from the global gene expression data of the heterogeneous samples (Gaujoux et al. 2011; Kuhn et al. 2011; Repsilber et al. 2010; Roy et al. 2006; Venet et al. 2001).

Figure 1 summarises the input data required by each type of method. It is clear from this figure that the type of deconvolution problems addressed by each method differ in their complexity, which increases as the input data requirement decreases:

1. Estimating proportions from known cell-type signatures: this is an overdetermined estimation problem separately for each sample, having generally more observations (genes) than coefficients to estimate (one proportion per cell type), which can potentially lead to very robust and accurate estimates (Abbas et al. 2009; Gong et al. 2011) (Figure 1a). The number of proportions to estimate is equal to the number of cell types r times the number of samples p ($r \times p$).
2. Estimating cell-type signatures from known proportions: this is also an overdetermined problem, separately for each gene, provided that there are more samples than estimated signatures. However, given the typical dimensions of gene expression data, with much more genes than samples, the estimation is expected to be less robust to measurement errors than in the previous case (Erkkila et al. 2010). The number of expression values to estimate is equal to the number of cell types r times the number of genes n ($r \times n$).
3. Estimating both signatures and proportions is a very loosely constrained problem, which requires more advanced techniques (Repsilber et al. 2010; Roy et al. 2006) or the incorporation of additional constraints, either technical (Venet et al. 2001) or driven by biological knowledge, as proposed by Gaujoux et al. (2011); Kuhn et al. (2011).

Strictly speaking, even complete deconvolution methods require some prior knowledge, often in the form of sets of marker genes, i.e. genes which are – essentially – expressed by a single cell type, among those expected to contribute to each sample's expression profile. These markers are used either to assign *a posteriori* estimated signatures to a given cell type (Repsilber et al. 2010), or *a priori* to enforce expected expression patterns to ensure biologically relevant signatures are recovered (Gaujoux et al. 2011). Kuhn et al. (2011) did not enforce marker gene patterns, but

computed the average expression of each set of markers, as proxies for the actual cell proportions. They then plugged these estimates into a standard linear regression analysis to correct for cell heterogeneity, and test for cell-specific differential expression between control and Huntington's disease human brain samples.



Graphic (b) was extracted from Shen-Orr et al. (2010), and modified to produce the other panels.

Figure 1: *Partial vs. Complete gene expression deconvolution.*
Partial deconvolution methods assume that either signatures (a) or proportions (b) are available and use them to infer the unknown proportions and signatures respectively. Complete deconvolution methods (c) infer both cell-type signatures and proportions from the global gene expression data.

1.1.1 General formulation

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 reasonable (Shen-Orr et al. 2010). Moreover, all approaches assume that gene expression at the cell level is independent of the cell type proportions, although this is expected not to be the case for some genes. Hence, both partial and complete gene expression deconvolution problems are most commonly formulated as linear models, with, for the complete problem, latent variables corresponding to cell type specific signatures (Abbas et al. 2009; Clarke et al. 2010;

Erkkila et al. 2010; Gong et al. 2011; Lähdesmäki et al. 2005; Lu et al. 2003; Repsilber et al. 2010; Shen-Orr et al. 2010; Venet et al. 2001; Wang et al. 2006). That is that the global expression value of gene i in sample j is the sum of its expressions in the r cell types:

$$x_i = \sum_{k=1}^{r} w_{ik} h_{kj} + \varepsilon,$$

where w_{ik} is the specific gene expression in cell type k, and h_{kj} the proportion of cell type k in sample j. Considering all genes together, this aggregates into the following approximate matrix decomposition problem:

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

Depending on the method, authors enforce additional constraints such as nonnegativity for gene expression values and proportions and sum to one constraints on proportions. As we have seen above, choosing a method to solve the problem in Equation (1) may depend on the availability of other auxiliary data, such as known signatures or cell type proportions. In the following we review some of the methods that have been proposed in common data settings.

1.1.2 Partial deconvolution methods

Partial deconvolution methods assume that either the signature matrix W or the proportion matrix H is available.

From signatures Most methods that estimate cell type proportions from quantitative cell type-specific signatures regress the global gene expression X on the known matrix W of cell-specific expression. Abbas et al. (2009) defined an optimised set of signatures for 17 different immune cell types, and proposed a heuristic algorithm based on standard linear regression to enforce nonnegative proportions, that are scaled to sum-up to one after fitting. We discuss the procedure used to build these signatures in more detail in Section 1.4. They apply their method to white blood cell samples from a cohort of Systemic Lupus Erythematosus (SLE) and healthy patients, and found differences in proportions that were corroborated by FACS and clinical observations (ibid.). Gong et al. (2011) proposed an alternative algorithm that incorporates the sum-up to one constraint on the proportions within the fitting process, aiming at improving their estimation. This seems indeed a natural constraint to impose, however, we found that these may make the algorithm sensitive to scale discrepancies between the signatures and the global expression profiles (see discussion in Section 2.2 and Figure 17). Although they used the same set of signatures, they applied their algorithm to a different dataset, making it difficult to compare its performance with the original method from Abbas et al. (2009). Lu et al. (2003) and Wang et al. (2006) used similar approaches to, respectively, identify dynamics of yeast cells at different stages of the cell cycle, and correct for varying proportions of mammary gland cell types in global mammary gene expression data. Clarke et al. (2010) proposed a different approach, not based on linear regression, for the case of two cell types, that requires signatures from only one of them.

From proportions Jacobsen et al. (2006) proposed a method to correct the expression values of cell-specific genes for sample heterogeneity, with application to the detection of their differential expression. Shen-Orr et al. (2010) proposed a method (*csSAM*) to perform differential analysis at the cell type levels, when proportions are available. Their approach consists in regressing the global gene expression on the proportions using standard linear regression, and incorporating the standard error estimates into a statistic that tests for differences in the estimated cell-specific expressions. Lähdesmäki et al. (2005) proposed an alternate least squares algorithm to optimise approximate estimates of proportions, as well as cross-validation and bootstrap methodologies to estimate the number of cell types and compute standard error estimates respectively. Erkkila et al. (2010) suggested a Bayesian method (*DSection*) for the same problem, which incorporates uncertainty in the initial proportions using Dirichlet prior distributions on the proportions. The

estimation is carried out by MCMC and Gibbs sampling. They showed that their method could correct noisy initial estimates of proportions to be more consistent with the observed expression data, and estimate standard errors that are less biased than those obtained by linear regression, making them more suitable for detecting cell-specific differential expression.

From marker genes Some genes, commonly called *marker genes*, are known to be – essentially – expressed only in a given specific cell type/tissue. Very recently, a few methods have been proposed to estimate cell type proportions based on the expression of marker genes. Miller et al. (2011) proposed a variety of aggregation method, including one based on gene co-expression networks, to extract/compute a single representative measure from all marker gene expression profiles of each cell type, which is used as an estimate of proportions. Bolen et al. (2011) calculated enrichment scores (Subramanian et al. 2005) of sets of marker genes in each individual sample expression profile, and used them as a proxy for cell type relative proportions, which was in turn used to predict the source of any given gene expression signature.

1.1.3 Complete deconvolution methods

Venet et al. (2001) pioneered the study of complete gene expression deconvolution, proposing an alternate nonnegative least-squares approach, using a heuristic to limit the correlations between the estimated cell type signatures. The same algorithm was subsequently used by Repsilber et al. (2010) and Lähdesmäki et al. (2005), although not using the same type of constraints. They applied their method to gene expression from normal and cancerous colon tissues, previously reported as being heterogeneous and containing varying proportions of muscle tissue (Alon et al. 1999). Using their method, they estimated a muscle tissue signature, whose corresponding proportion matched a muscle enrichment index that had been proved to correlate well with the samples' muscle content.

Repsilber et al. (2010) proposed an Nonnegative Matrix Factorization (NMF) algorithm (*deconf*) that corresponds exactly to Venet et al.'s algorithm, dropping the correlation constraints. They evaluated their method using simulated data, in terms of power of detection of differential expression, sensibility to sample size and normalisation method, and proposed a strategy, based on random forest (Breiman 2001), to apply the method as a classifier. In (Gaujoux et al. 2011) we proposed another NMF algorithm, designed to estimate more meaningful signatures, which reflect prior knowledge of marker genes. Kuhn et al. (2011) used marker genes to first estimate a proxy for each cell type proportions

The formulation of gene expression deconvolution as in Equation (1), in addition to the physical nonnegativity constraints makes the NMF theoretical framework a natural choice for developing complete deconvolution algorithms. In fact, some authors did explicitly label their recent approach as NMF methods (Gaujoux et al. 2011; Repsilber et al. 2010). Even the algorithms proposed by Venet et al. or Lähdesmäki et al. are typical NMF algorithms, although not named as such, and use a classical alternate least-squares strategy.

1.2 Accuracy of deconvolution methods

The performance of computational deconvolution methods is generally assessed on datasets, for which the quantities to be determined, i.e. cell signatures or proportions, are known and used as a gold standard. These are typically data generated from titration experiments, where samples are artificial mixtures of pure cell types, or from experiments where physical separation has been performed for each sample. The comparison is commonly made in term of accuracy/precision (i.e. mean difference/variance) or Pearson correlation between the true and estimated quantities. We report here below the performances achieved by some of the deconvolution methods reviewed in the previous sections.

Using cell type signatures defined from pure samples, Abbas et al. (2009) achieved good accuracy (bias=2.4±1.4%) and precision (s.d. 0.78±0.52%) on a controlled mixture experiment of immune cell lines, and a mean bias of 1.3 % when estimating the proportion of three T-cell subsets in PBMC samples. When applied to blood samples from SLE and healthy patients, their method a

resonably good Pearson correlation of 0.5196 with measured CBC data, and could identify differences between cases and controls in the proportions of several immune subsets. The aggregated measures proposed by Miller et al. (2011) correlated well with actual proportions, when applied to both the controlled mixture experiment from Abbas et al. (2009) ($0.82 \geq r \leq 0.99$), and real clinical blood samples from Grigoryev et al. (2010), to a lower extend however ($0.4 \geq r \leq 0.6$ in most cases). Bolen et al. (2011) applied their enrichment score approach to 161 PBMC samples collected from different studies, which showed relatively good correlations to proportions estimated using the partial deconvolution method from Abbas et al. (2009) ($0.65 \geq r \leq 0.87$ for most cell types, but only $r = 0.21$ for T-cells, which they attributed to the quality of the markers used for this particular cell type). We proposed ourselves a semi-supervised NMF approach (Gaujoux et al. 2011), which significantly improved Pearson correlations between true and estimated cell type proportions from about 0.5 to approximately 0.8, when compared to unsupervised NMF methods, using the controlled mixture dataset from Abbas et al. (2009). Finally Kuhn et al. (2011)'s marker-based linear regression reconstructed cell-specific expression levels from a controlled mixture of pure brain cell expression profiles with Pearson correlations between 0.92-1.00. Moreover, they showed that using few markers is sufficient for computing an efficient proxy for cell proportions, from which moderate expression change can be detected, provided that sample-to-sample variations account for most of the total variance (compared to measurement noise). For instance, on simulated data, 5 markers were enough to detect a 1.5 log fold change, while using more markers increased power only marginally.

1.3 Benchmark gene expression datasets

For deconvolution algorithms to be developed, tested and validated, it is critical to be able to benchmark and compare their performances on real data. For this purpose, gene expression repositories such as the Gene Expression Omnibus (GEO) database (Barrett et al. 2010) or ArrayExpress (Parkinson et al. 2009) are extremely useful data sources. In the context of deconvolution, experiments that contain expression data from pure cell/tissue types, or from mixed samples with known controlled/measured mixture proportions are especially interesting. These provide robust ground truth references to calibrate and test deconvolution algorithms. As a matter of fact, several such datasets are publicly available, and have been used in this way in recent publications (Erkkila et al. 2010; Gaujoux et al. 2011; Gong et al. 2011; Miller et al. 2011; Shen-Orr et al. 2010).

1.3.1 The sample annotation issue

These datasets are stored in a standardised format, that can be can be conveniently accessed programmatically, e.g in R using the packages *GEOquery* package[1] (Davis et al. 2007) or *ArrayExpress* package[2] (Kauffmann et al. 2009). These packages can download datasets given their accession numbers in GEO and ArrayExpress respectively, and load them in a format that integrates well with other Bioconductor analysis routines. However, although being MIAME compliant (Brazma et al. 2001), associated sample phenotypic annotation data, such as experimental factors and in particular known cell proportions when available, are often encoded in annotation fields by the submitting users, with no fixed standard. ArrayExpress is more flexible than GEO with respect to the storage of such annotations, as multiple experimental factors and sample attributes are appropriately implemented in separate named data fields, with identified levels (i.e. the set of possible values for a given factor). Multiple sample annotations also have separate dedicated data fields in GEO, but these are generically named *characteristics_ch1*, *characteristics_ch1.1*, etc. and have no defined levels. The main issue however arises from the fact that annotations are often stored all into a single composite data field, possibly not even the appropriate one, e.g. in the fields normally dedicated to sample or source names. Hence, even if the data structure exists, neither GEO nor ArrayExpress can automatically extract or infer all meaningful sample annotations from composite fields or fields other than the dedicated ones. Figures 2 and 3 illustrate these issues

[1] http://www.bioconductor.org/packages/release/bioc/html/GEOquery.html
[2] http://www.bioconductor.org/packages/release/bioc/html/ArrayExpress.html

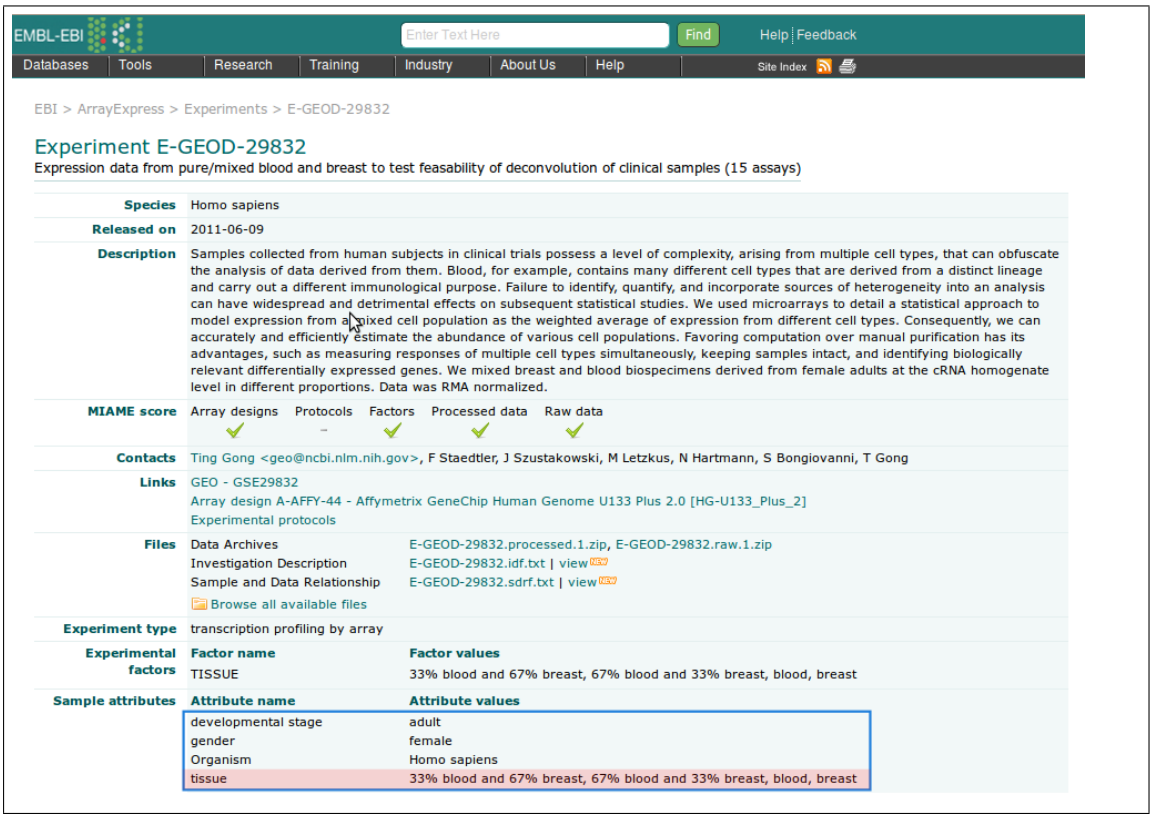


Figure 2: *Screenshot of the description page for dataset* E-GEOD-29832 *on ArrayExpress. Multiple sample attributes are correctly stored into separate data fields (blue frame), but mixture proportions are encoded into a single composite field (red highlighted text).*

for datasets E-GEOD-19830[3], GSE19830[4] and GSE5350[5], where the cell proportions have to be extracted from composite data fields encoded in various different ways.

Relevant phenotypic annotations are also frequently available as separate supplementary data, which can take a variety of more or less convenient formats (e.g. txt, csv, xls, doc, pdf files), possibly being ordered or identified with sample IDs in a different way than in the repository expression data. Other times the data is not directly available, but must be enquired of the authors. Overall, this means that these data are often not as straightforwardly usable as they could be, requiring some pre-processing in order to get, extract or split relevant phenotypic data into a more meaningful format. Researchers working with deconvolution methods would certainly benefit from easy access to suitably pre-formated benchmark datasets, which would alleviate pre-processing, and allow direct usage of data.

1.4 Cell/Tissue-specific quantitative signatures

Some partial gene expression deconvolution methods require the availability of a basis matrix, i.e. cell type-specific expression signatures. These are typically used within the standard linear regression framework, to estimate cell proportions in complex samples (Abbas et al. 2009; Gong et al. 2011; Lu et al. 2003; Wang et al. 2006). Because there are usually much more probesets in the basis signatures than cell type proportions to estimate, these methods are potentially very robust and accurate. Even for the methods that do estimate the signatures (Erkkila et al. 2010;

[3] http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-19830
[4] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19830
[5] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5350

```
# Dataset: GSE19830 [Shen-Orr et al. (2010)]
GSE19830 <- getGEO("GSE19830")[[1]]
levels(GSE19830$characteristics_ch1)
```

```
##  [1] "tissue: 0 % Liver / 0 % Brain / 100 % Lung"
##  [2] "tissue: 0 % Liver / 100 % Brain / 0 % Lung"
##  [3] "tissue: 100 % Liver / 0 % Brain / 0 % Lung"
##  [4] "tissue: 25 % Liver / 70 % Brain / 5 % Lung"
##  [5] "tissue: 45 % Liver / 45 % Brain / 10 % Lung"
##  [6] "tissue: 50 % Liver / 30 % Brain / 20 % Lung"
##  [7] "tissue: 50 % Liver / 40 % Brain / 10 % Lung"
##  [8] "tissue: 55 % Liver / 20 % Brain / 25 % Lung"
##  [9] "tissue: 55 % Liver / 30 % Brain / 15 % Lung"
## [10] "tissue: 5 % Liver / 25 % Brain / 70 % Lung"
## [11] "tissue: 60 % Liver / 35 % Brain / 5 % Lung"
## [12] "tissue: 65 % Liver / 34 % Brain / 1 % Lung"
## [13] "tissue: 70 % Liver / 25 % Brain / 5 % Lung"
## [14] "tissue: 70 % Liver / 5 % Brain / 25 % Lung"
```

```
# Dataset: GSE5350 (MACQ project) [HUMAN]
GSE5350 <- getGEO("GSE5350")[[1]]
# extract Affy Human dataset (i.e. GEO platform GPL570)
i <- grep("GPL570", attr(GSE5350, "names"))
GSE5350 <- GSE5350[[i]]
# extract only mixture experiments
GSE5350[, grep("^MAQC", GSE5350$source_name_ch1)]
levels(droplevels(GSE5350$source_name_ch1))
```

```
## [1] "MAQC sample A, i.e., Stratagene Universal Human Reference RNA (UHRR, Catalog #740000)"
## [2] "MAQC sample B, i.e., Ambion Human Brain Reference RNA (HBRR, Catalog #6050)"
## [3] "MAQC sample C, i.e., MAQC samples A and B mixed at 75%:25% ratio (A:B)."
## [4] "MAQC sample D, i.e., MAQC samples A and B mixed at 25%:75% ratio (A:B)."
```

Figure 3: *Examples of encoded phenotypic data from GEO data fields* characteristics_ch1 *or* source_name_ch1 *for datasets* `GSE19830` *and* `GSE29832` *respectively. Although the interesting data are present, the fields require pre-processing in order to extract the data of interest, the cell type proportions in this case.*

Repsilber et al. 2010; Shen-Orr et al. 2010), having reference basis signatures to compare with makes it possible to assign the estimated signatures and/or assess their meaningfulness.

In many studies, the basis signatures are obtained from the complex samples themselves, by separating, purifying and assaying the constituting cell types of interest. However, one would ideally want signatures that are defined and optimised once for a given set of cell types, and may be used for partial deconvolution of expression data from other independent experiments, without the need for physical separation. For this approach to be valid, such generic signatures must clearly be robust with regards to a variety of experimental factors. Moreover, an efficient probe identifier cross-platform conversion pipeline should be designed, so that gene expression data generated on a given platform can be deconvolved using signatures defined on another.

1.4.1 The delicate choice of basis signatures

When defining signatures for murine mammary gland development stages, Wang et al. (2006) left aside cell type specific genes that were highly regulated, as their expression could prove to be too sensitive to variations in biological conditions, and hamper the deconvolution accuracy. They first selected tissue type specific probesets using a combination of pairwise comparisons between all tissues (based on t-tests) and step-wise discriminant analysis, and computed the mean expression within each tissue, for each time point in their time-course experiment. Probesets whose expression did not correlate well with their associated tissue mean expression (Pearson correlation coefficient less than 0.5), were then removed from the signatures.

In order to design an efficient basis matrix for whole blood deconvolution, Abbas et al. (2009) considered the correlations between signatures and their collective ability to accurately deconvolve global gene expression. Abbas et al. (2005) compiled microarray gene expression data for six key human immune cell types and their activated and differentiated states. They used these data to create the curated database Immune Response In Silico (IRIS) [6], that is composed of genes selected for their specific expression in immune cells. Based on the same expression data, Abbas et al. (2009) developed a partial deconvolution method for estimating the proportions of 17 immune cell-types from microarray expression data from real blood samples, using only an optimised subset of genes. They observed that the accuracy of the proportion estimates was closely correlated with the condition number of the basis matrix: the lower the condition number the better the goodness of fit, suggesting its use as a predictive measure of the deconvolution power of a given set of signatures (see Section 4). After selecting and ranking the probesets according to their cell type specificity (using a t-test approach), they built a sequence of basis matrices, by including an increasing number of top probesets, and chose the one with the minimum condition number (359 probesets).

The deconvolution basis matrix designed by Abbas et al. (ibid.) proved to perform well on real clinical data. When used to deconvolve white blood cell samples from a cohort of 72 SLE and 45 healthy patients, the computed proportions of Lymphocytes, Monocytes and Neutrophils showed an average Pearson coefficient of 0.5196 with the corresponding measured abundance by CBC (ibid.). Using the same set of signatures on whole-blood samples from 28 Multiple Sclerosis (MS) patients and 10 healthy donors, Gong et al. (2011) achieved Pearson correlation coefficients of 0.85, 0.62 and 0.61 for each of these cell types respectively. Although the optimisation technique used in this case was different from the one used by Abbas et al. (2009), it still fundamentally relies on the robustness of the set of basis signatures. Having such data integrated with appropriate partial deconvolution methods, would therefore enable the quick and detailed deconvolution of blood gene expression data.

1.5 Marker gene lists

To be applicable, some complete deconvolution methods requires a list of marker genes, known to be specifically expressed in different cell types or tissues. They used these marker genes *a posteriori* to map the estimated signatures to known cell types (Repsilber et al. 2010) or as prior

[6] http://share.gene.com/clark.iris.2004/iris/iris.html

knowledge to enforce expression patterns on *de facto* assigned signatures, as in the semi-supervised method proposed in Gaujoux et al. (2011). Known cell/tissue marker lists may also be used in classical enrichment analysis to assess, the pertinence or biological significance of given expression profiles. Bolen et al. (2011) used PBMC subset-specific genes for predicting the most likely cellular source of a predefined gene list from global PBMC expression data. The query gene list would typically consist of a disease gene signature, as generated by gene expression differential analysis of case vs. control patients in infectious disease studies. Using GSEA (Subramanian et al. 2005), they computed, for each sample, the enrichment score of the query gene list and each cell subset-specific gene set – sorted by expression values. The cellular source of the disease signature was predicted to be the cell subset whose scores are the most correlated with the scores obtained for the query list. Finally, as already mentionned, Kuhn et al. (2011) used maker genes to directly estimate cell proportions, in a pre-step of their linear regression approach.

1.5.1 Public databases

A number of databases have been explicitly designed to provide cell/tissue specific gene level information, in a tissue-centred manner. We mentioned above the IRIS database, that focuses on immune cell types, providing marker genes[7] for B cells, dendritic cells, monocytes, neutrophils, Natural Killer (NK) cells and T cells, as well as for the lymphoid and myeloid lineages, and the general category of all these immune cells. TiGER (Liu et al. 2008a) stores tissue-specific expression profiles for UniGene clusters in 30 human tissues based on UniGene EST database (Boguski et al. 1993), with a focus on Transcription Factor (TF) interaction and cis-regulatory modules. TissueDistributionDB[8] (Kogenaru et al. 2009) is a repository of tissue-distribution profiles for 20 model organisms. It combines tissue expression profiles from UniGene with the *Tissue Ontology* available from BRENDA[9] (Gremse et al. 2011), to standardise tissue information and compute tissue specificity measures for each gene at four different anatomical levels of the ontology, from the more general (e.g. hematopoietic system) to the more specific localisation (e.g. blood platelet). TiSGeD[10] (Xiao et al. 2010) provides information on tissue-specific genes, compiled from multiple public microarray datasets (over 123 000 expression profiles from 107 human tissues, 67 mouse tissues and 30 rat tissues), and notably links to relevant literature. C-It[11] (Gellert et al. 2010) is a database of tissue-enriched genes from human, mouse, rat, chicken and zebrafish. Its specific objective is to integrate tissue information from UniGene from multiple organisms, together with microarray and SAGE data, as well as using specific methods for handling alternative splicing from exon array data. The aim is to provide a more comprehensive view of transcriptional profiles, and reduce the number of false positive, given that evolutionary conservation of tissue specificity provides greater confidence. Although its main focus is on identifying uncharacterised genes, well characterised genes may be retrieved using loose filtering criteria on the number of associated publications and Medical Subject Headings (MeSH) terms. Finally, VeryGene[12] (Yang et al. 2011) is a database for the annotation of human tissue-specific genes, with a primary focus on integration with disease association and drug targets. At the time of writing, it contains entries for 3960 genes covering 128 normal tissue/cell types compiled from the expression profiling of two large microarray datasets (around 4,000 samples combined) (Liang et al. 2006; Su et al. 2004).

The heterogeneity in format, accessibility and biological identifiers across these databases limits the potential of their data. Programmatic access to their complete data, together with a common data structure, would make their integration and use with different gene expression datasets easier, whether it is for gene expression deconvolution or enrichment analysis.

[7] More precisely Affymetrix probesets
[8] http://genius.embnet.dkfz-heidelberg.de/menu/tissue_db
[9] http://www.brenda-enzymes.org/ (Organism-related information > Source Tissue)
[10] http://bioinf.xmu.edu.cn/databases/TiSGeD
[11] http://c-it.mpi-bn.mpg.de
[12] http://www.verygene.com

1.6 Availability of deconvolution algorithms

Gene expression deconvolution receives constant interest in bioinformatics research, and new methodologies are regularly published. We reviewed in Section 1.1 the different deconvolution methods of which we are aware. Methods and algorithms related to gene expression deconvolution are often developed in *R*, and sometimes available as *R* packages, under relatively liberal license terms. From the point of view of access to the algorithms and free open-source software development, this is actually not too bad a situation.

Abbas et al. (2009) briefly described the main *R* functions they used in the implementation of their partial deconvolution approach. The *csSAM* algorithm[13] from Shen-Orr et al. (2010) and the semi-supervised approach[14] from Gaujoux et al. (2011) are available as *R* packages, although not on CRAN. The univariate strategy for predicting cell type abundances based on a single representative measurement of Miller et al. (2011) is available as part of the *WGCNA* package[15] (Langfelder et al. 2008). The cell subset prediction algorithm from Bolen et al. (2011) based on the Gene Set Enrichment Analysis (GSEA) algorithm (Subramanian et al. 2005) is implemented via a web interface, that uses *R* to run the analysis at the backend. The *DSection* algorithm and the quadratic programming approach from Gong et al. (2011), however, are implemented in Matlab®, with the complete code directly available for *DSection* only.

Given the variety of methods and implementations (i.e. of user interfaces), a standardised and unified interface for running easily a variety of deconvolution methods, in most common data settings, would be very useful. This would help popularise computational deconvolution, an inexpensive technique that can provide insights into underlying biological processes, at the cell type level.

In order to facilitate the application and development of gene expression deconvolution methods, we developed an *R* package called *CellMix*, the principal objectives of which are to provide *a*) easy access to real benchmark data, and especially marker gene lists; *b*) implementations of some common methods; *c*) utilities for assessing results and developing new methods.

2 Results

The *CellMix* package builds upon Bioconductor (Gentleman et al. 2004) and the *NMF* package (Gaujoux et al. 2010), to provide a flexible general framework for gene expression deconvolution methods. It aims at alleviating the challenges reviewed in the introduction of this vignette, by defining a rich programming interface around three internal extensible registries dedicated to benchmark datasets, marker gene lists and deconvolution methods respectively. This section describes the main features of the *CellMix* package, and illustrates its capability with some concrete examples. More technical details on the implementation itself can be found in Section 3.

2.1 Benchmark expression data

We designed a small internal repository of 12 datasets compiled from a variety of published studies on cell/tissue specific gene expression or deconvolution methods. We typically looked for data that contain cell type specific signatures or samples with mixture proportions, and ideally both. We selected in particular data that have been used by – multiple – authors for validating deconvolution approaches, such as datasets GSE11058[16] and GSE19830[17] from Abbas et al. (2009) and Shen-Orr et al. (2010) respectively, also used by Gaujoux et al. (2011); Gong et al. (2011); Miller et al. (2011).

Table 1a summarises the list of selected datasets together with some of their characteristics (size, number of cell/tissue type). Columns *Basis* and *Coef* indicate with a ✓which compositional

[13] http://buttelab.stanford.edu/public:data
[14] http://web.cbio.uct.ac.za/~renaud/paper/meegid-deconvolution
[15] http://cran.r-project.org/package=WGCNA
[16] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11058
[17] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19830

data is available for each dataset – a "-" meaning that the data is not available in the registry. Since the datasets are already stored and available on public repositories, it would be inefficient to duplicate the data storage itself. Hence for each dataset, we defined a set of functions that together compose a pre-processing pipeline, that is applied to the original – normalised – data as available in public repositories. The pipeline combines the global expression data and, when available, its constituting cell type specific signatures and proportions, into a single data object (see Section 3.2). These data objects are directly usable for benchmarking deconvolution methods, and implement the standard interfaces defined by the Bioconductor *Biobase* package[18] and the *NMF* package.

For a given dataset, the complete pipeline can be performed in a single call, as illustrated in Figure 4, which shows sample code for loading dataset *GSE11058*. When loaded for the first time, the data is downloaded from GEO, and stored locally, allowing for faster subsequent access, i.e. in a different R sessions. Moreover, sample annotations are directly accessible, dedicated methods enable to extract expression data for pure or mixed samples only, or access the known mixture proportion or cell type signatures when available.

2.2 Whole blood deconvolution

The *CellMix* package includes the deconvolution basis matrix from Abbas et al. (2009), and provides a dedicated interface function (`gedBlood`) to easily estimate the proportions of 17 immune cell types from whole blood or white blood cell gene expression data. The interface also facilitates the aggregation of these detailed proportions into proportions of the more general cell type categories lymphocytes, monocytes, neutrophils. These aggregated proportions may be assessed by comparing them with actual CBC data when available, as it frequently is the case in recent clinical studies. Figure 5 shows sample code that illustrates how such proportions can be estimated straightforwardly for dataset GSE20300[19], which contains gene expression data of whole blood samples from stable and acute rejection pediatric kidney transplant, for which CBC data are available. These data where used by Shen-Orr et al. (2010) to illustrate the use of the *csSAM* algorithm, and to identify differential expression between the two clinical groups of samples at the cell type level, by deconvolving their global gene expression based on the measured CBC proportions. We use it here the other way round, since we estimate proportions from independently defined cell-specific signatures. Because this dataset is included in *CellMix* internal data registry, the CBC data are readily accessible and used transparently by the different specialised functions.

By default, the proportions are estimated using the standard linear regression approach from Abbas et al. (2009). Alternative estimation methods such as the quadratic programming approach from Gong et al. (2011) may also be used (see Section 3.4). Moreover, the signatures are log-transformed – if the global expression is itself in log-scale – and quantile-normalised together with the global expression data prior to estimation (Bolstad et al. 2003). Our experiments show that these pre-processing steps affect the accuracy and or correlation of the estimations as shown in Figure 17, suggesting that scaling discrepancies must be taken into account in this kind of analysis.

The scatter plot shows that the estimated proportions are remarkably well correlated with the actual measured proportions, to a lower extent for monocytes. These results are within the range of correlations reported by other authors using the same set of signatures on different datasets (Abbas et al. 2009; Gong et al. 2011), but achieve better separate correlations, probably due to the joint normalisation of the signatures and the target matrix. To our knowledge, this is the first time these data have been used in this way. The ease with which the results above are generated highlights the usefulness of the *CellMix* package.

[18] http://www.bioconductor.org/packages/release/bioc/html/Biobase.html
[19] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20300

```
## load data
emix <- ExpressionMix("GSE11058")

## get relevant sample annotation cell type
emix$Type

##  [1] Jurkat Jurkat Jurkat IM-9   IM-9   IM-9   Raji   Raji   Raji   THP-1
## [11] THP-1  THP-1  MixA   MixA   MixA   MixB   MixB   MixB   MixC   MixC
## [21] MixC   MixD   MixD   MixD
## Levels: IM-9 Jurkat MixA MixB MixC MixD Raji THP-1

# cell type of origin
emix$CType

##  [1] T        T        T        B        B        B        B
##  [8] B        B        Monocyte Monocyte Monocyte MixA     MixA
## [15] MixA     MixB     MixB     MixB     MixC     MixC     MixC
## [22] MixD     MixD     MixD
## Levels: B MixA MixB MixC MixD Monocyte T


## get the reference mixing proportions
coef(emix)[, c(1, 4, 7, 13:15)]

##        GSM279589 GSM279592 GSM279595 GSM279601 GSM279602 GSM279603
## Jurkat         1         0         0     0.250     0.250     0.250
## IM-9           0         1         0     0.125     0.125     0.125
## Raji           0         0         1     0.250     0.250     0.250
## THP-1          0         0         0     0.375     0.375     0.375

## get the reference cell-type specific signatures
head(basis(emix), 3)

##           Jurkat IM-9   Raji  THP-1
## 1007_s_at  341.7  339 1444.0  209.6
## 1053_at   1997.5 2377  903.1 1488.6
## 117_at     153.6  325  173.6  201.9
```

Figure 4: *Sample code for loading gene expression deconvolution datasets. The loaded data object combines both the global gene expression and composition data. It also contains relevant sample annotations, extracted from the original composite annotation sources.*

```r
# load data
e <- ExpressionMix("GSE20300")
# compute proportions (show processing)
res <- gedBlood(e, verbose = TRUE)

## Loading basis signature from Abbas et al. (2009) ... OK [359 features x 17 cell types]
## Mapping signature ids onto target ids (method: auto) ... OK [322 features x 17 cell types]
## Limit/reorder to common set of features ... OK [322 features x 17 cell types]
## Checking data dimension compatibility ... OK [322 features x 17 cell types]
## Using features:
##  * Data: '232165_at', '201369_s_at', ..., '223809_at'
##  * Signatures: '232165_at', '201369_s_at', ..., '223809_at'
## Checking log-scale ... data:YES - signatures:NO
## Applying log-transform to signatures (base 2) ... OK
## Normalizing signatures and target together (method: quantiles) ... OK
##  Using ged algorithm: "lsfit"

## NMF algorithm:  'lsfit'

##   Estimating cell proportions from cell-specific signatures [lsfit]
##  Timing:
##    user  system elapsed
##   0.768   0.028   0.799
##  GED final wrap up ...  OK

# aggregate into CBC
cbc <- asCBC(res)
# plot against actual CBC
## profplot(e, cbc)
```

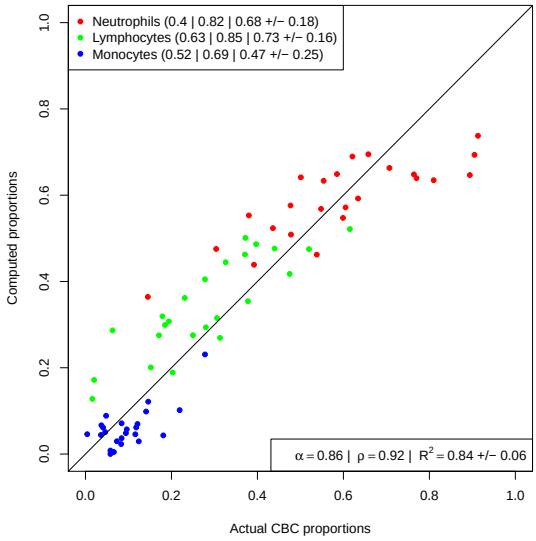


Figure 5: *Computed proportions vs. Actual CBC proportions of lymphocytes, monocytes and neutrophils for dataset* GSE20300. *The deconvolution is performed using the* `qprog` *algorithm in combination with the basis signature matrix from Abbas et al. (2009). Computed proportions are obtained from the aggregation of proportions of the detailed subset of immune cells into their respective category. First diagonal is shown as a reference for perfect agreement.*

```
m <- MarkerList("Palmer")
summary(m)

## <object of class MarkerList>
## Types: 5 ['B', 'CD8', ..., 'T']
## Mode: numeric
## Markers: 907
## IDtype: unknown ['FREB', 'CD20', ..., 'LRIG1']
## Values: [0.953208, 0.950334, ..., 0.517031]
## Source: org.Hs.eg.db
## Breakdown:
##      B    CD8  GRANS LYMPHS      T
##    314     22    344     39    188
```

Figure 6: *Sample code for loading a marker gene list from the internal registry.*

2.3 Marker gene lists

The *CellMix* package includes 8 lists of marker probesets defined from public tissue-centred databases, as well as from three gene expression microarray studies of PBMC samples. These latter lists are valuable resources for the deconvolution of clinical human blood samples. They include the refined subset of immune genes that compose the basis deconvolution matrix from Abbas et al. (2009), data from Palmer et al. (2006), who defined markers for B-cells, CD4+ T-cells, CD8+ T-cells, lymphocytes and granulocytes using cDNA microarrays from purified cells, and data from HaemAtlas as generated by Watkins et al. (2009), who more recently defined markers for CD4+ T-cells and CD8+ T-cells, lymphocytes, CD14+ monocytes, CD19+ B-cells, CD56+ NK cells, and CD66b+ granulocytes, as well as for Erythroblasts (EBs) and Megakaryocytes (MKs), using Illumina BeadChip arrays.

Similarly to the benchmark expression data, the marker lists are organised in an internal registry, that facilitates their management and the addition of new lists, as they become publicly available. For completeness and reproducibility, all the functions used to generate them from their respective original data files also available and documented in the package. Beside the markers' information itself (i.e. identifiers and possibly specificity scores), the registry contains important auxiliary data, such as the relevant annotation package that is required in practice for some marker lists to be used in combination with expression data, generated on different platforms (see example below). We also defined a specific data structure that provides a rich interface to create and manipulate these data in a convenient way, well integrated with R and Bioconductor standards. Table 1b summarises the marker lists currently available in the *CellMix* package along with their respective auxiliary data.

2.3.1 Specificity scores

Marker lists are stored as complete as possible, and each marker is associated with a specificity score (e.g. percentage expression, sparseness) when such value is available or can be computed. We designed the registry interface to make it possible to filter the lists based on these scores when retrieving the data from the registry, the default being to use some stringent threshold. Figure 6 shows sample code that illustrates how maker lists are loaded from the registry, and display the summary view of the returned data structure. The marker list shown was defined by Palmer et al. (2006), for which the specificity scores (0.953208, 0.950334, etc.) corresponds to the correlations of each marker's expression profile with the theoretical relative abundance profile of their assigned cell type as provided by Palmer et al.

The list of markers derived from the `Abbas` dataset provides another example of specificity score, that we defined as the sparseness each marker's expression profile, as defined by Hoyer

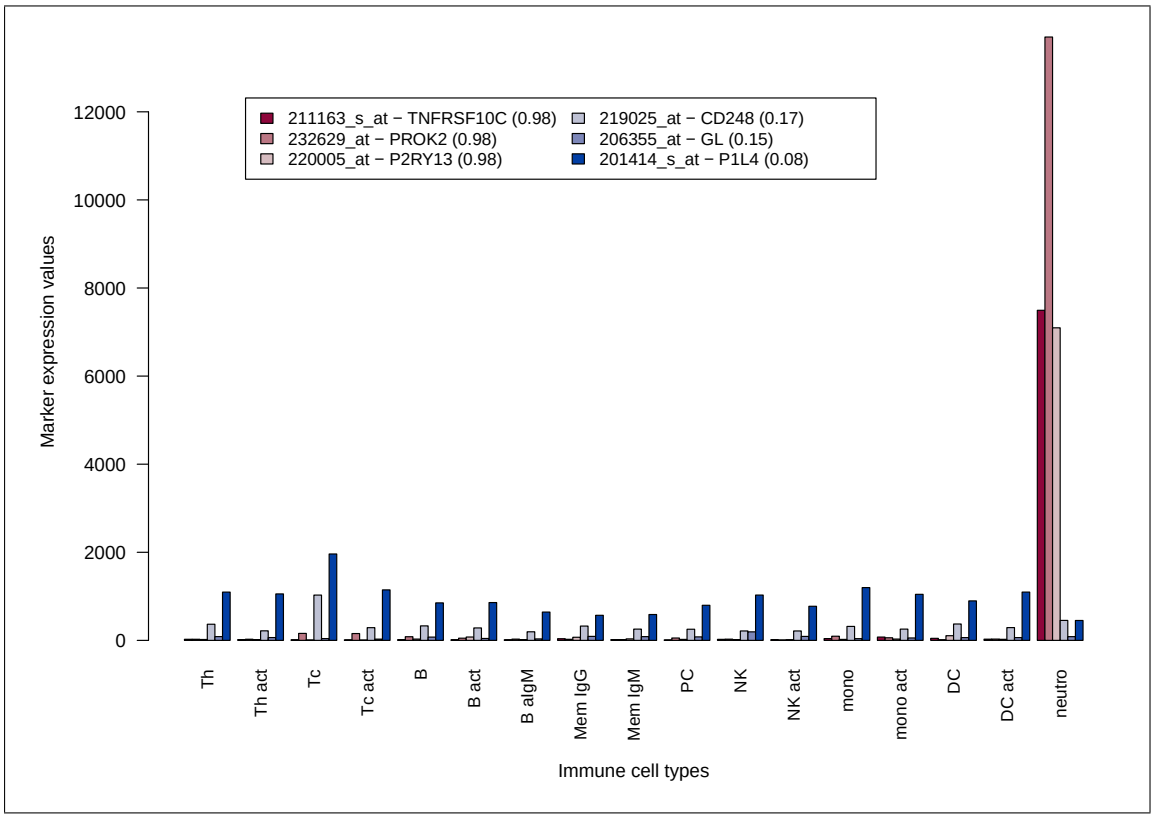


Figure 7: *Expression profiles of six probesets with low/high sparseness from the* `Abbas` *dataset. Bars are coloured according to the sparseness value: shades of red (resp. blue) for high (resp. low) sparseness. Numbers in parenthesis show the sparseness values of each probeset.*

(2004):

$$\text{sparseness}(x) = \frac{\sqrt{n} - \frac{\sum \|x_i\|}{\sqrt{\sum x_i^2}}}{\sqrt{n} - 1},$$

for any vector x in $\mathbb{R}^n$, i.e., here the vector of expression values of a given marker across all samples.

The default is to extract only the probesets whose sparseness is greater than 0.8, which filters out probesets that are expressed by multiple cell types, but the complete list may be retrieved using a null sparseness threshold. The bar chart in Figure 7 illustrates the effect of this filtering step, by showing the expression profiles of six probesets across the immune cell samples from this dataset, chosen from both ends of the complete sparseness spectrum. The bars are coloured from dark red to dark blue by decreasing sparseness, shown in parenthesis in the legend. Probesets with low sparseness values (blue) are expressed relatively uniformly over almost all the cell types, while those with high sparseness values (red) are highly specific, in this case all to neutrophils. As a result, by default, the red probesets would be included in the marker list, while the blue would be excluded.

2.3.2 Using marker lists across platforms

In practice, gene expression data can be generated in multiple ways, using different technologies (e.g. microarrays, next generation sequencing) provided by different manufacturers (e.g. Affymetrix, Illumina, Roche). When working with microarray data in particular, one has to face the well known issue of probe annotation, which becomes all the more critical if one wants to integrate data from multiple platforms (Carter et al. 2005; Draghici et al. 2006; Mecham et al.

(a)

	Description	Features	Samples	Types	Mixed	Pure	Basis	Coef	Annotation	Reference
GSE29832	Pure/mixed blood and breast to test deconvolution of clinical samples	54675	15	2	✓	✓	✓	✓	hgu133plus2.db	Gong et al. (2011)
GSE24223	Deconvoluting Early Post-Transplant Immunity Using Purified Cell Subsets	54675	179	5	✓	✓	-	-	hgu133plus2.db	Grigoryev et al. (2010)
GSE19830	Pure/mixed brain, liver and lung to test statistical deconvolution [Rat]	31099	42	3	✓	✓	✓	✓	rat2302.db	Shen-Orr et al. (2010)
GSE20300	Whole blood from stable and acute rejection pediatric kidney transplant	54675	24	5	✓	-	-	✓	hgu133plus2.db	Shen-Orr et al. (2010)
GSE5350	MicroArray Quality Control (MAQC) Project: Affymetrix HG-U133 Plus 2.0	54675	120	2	✓	✓	✓	✓	hgu133plus2.db	Shi et al. (2006)
GSE11057	Memory T Cell Subsets: Central memory, Effector memory, Naive	54675	17	3	✓	✓	✓	-	hgu133plus2.db	Abbas et al. (2009)
GSE11058	Immune Cell Line Mixtures: Jurkat, IM-9, Raji, THP-1	54675	24	4	✓	✓	✓	✓	hgu133plus2.db	Abbas et al. (2009)
GSE22886_A	IRIS: Resting and activated human immune cells [HG-U133A]	22283	114	11	-	✓	-	-	hgu133a.db	Abbas et al. (2005)
GSE22886_B	IRIS: Resting and activated human immune cells [HG-U133B]	22645	114	11	-	✓	-	-	hgu133b.db	Abbas et al. (2005)
GSE33076	Linearity of amplification between gene expression values and mRNA in retina cells	22347	24	2	✓	✓	✓	✓	mogene10stprobeset.db	Siegert et al. (2012)
GSE3649	Individuality and variation in gene expression patterns in human blood	36794	70	5	✓	-	-	✓		Whitney et al. (2003)
E-TABM-633	The HaemAtlas: Transcription profiling of differentiated human blood cells	46693	50	8	-	✓	✓	✓	illuminaHumanv2.db	Watkins et al. (2009)

(b)

	Description	Organism	Types	Markers	idType	Annotation	Reference
IRIS	Immune Response In Silico: B, T, NK and dendritic cells, monocytes and neutrophils	Human	9	2270	.Affymetrix	hgu133a.db, hgu133b.db	Abbas et al. (2005)
Abbas	Optimised set of immune genes for deconvolution of blood samples	Human	12	122	.Affymetrix	hgu133a.db, hgu133b.db	Abbas et al. (2009)
TDDB_HS	TissueDistributionDB: UniGene EST distribution profiles using Tissue Ontology from BRENDA	Human	65	14732	UNIGENE	org.Hs.eg.db	Kogenaru et al. (2009)
TDDB_RN	TissueDistributionDB: UniGene EST distribution profiles using Tissue Ontology from BRENDA	Rat	19	3915	UNIGENE	org.Rn.eg.db	Kogenaru et al. (2009)
Palmer	Markers for B-cells, CD4+ and CD8+ T-cells, lymphocytes and granulocytes	Human	5	907	SYMBOL	org.Hs.eg.db	Palmer et al. (2006)
HaemAtlas	HaemAtlas markers for CD4+ and CD8+ T-cells, lymphocytes, monocytes, B-cells, NK cells, and granulocytes	Human	8	2069	.Illumina	illuminaHumanv2.db	Watkins et al. (2009)
TIGER	TiGER database: based on UniGene EST distribution profiles	Human	30	7743	UNIGENE	org.Hs.eg.db	Liu et al. (2008b)
VeryGene	VeryGene database: based on two large microarray datasets	Human	127	10102	ENTREZID	org.Hs.eg.db	Yang et al. (2011)

Table 1: *Data available in the CellMix package: (a) Benchmark and signature expression data, (b) lists of tissue-specific marker genes/probesets.*

2004). This issue is relevant in the context of deconvolution, as the global or tissue/cell specific expression data with which the marker gene lists are to be used may come from a different source, e.g. markers provided as UniGene or Affymetrix IDs with data generated on an Illumina chip. One may also want to compare marker lists generated for different organisms (hence different platforms), which would provide insights into the conservation of tissue/cell specific expression patterns or help build more refined and robust marker lists.

The marker lists available in the *CellMix* package were generated from diverse microarray datasets and/or the UniGene EST database. In order to keep as much information as possible from the original data source, the markers are stored using their original identifiers, which can be manufacturer probeset ids (e.g. '123456_at', 'ILMN_123456'), UniGene ids (e.g. 'Hs.123456', 'Rn.123456'), Entrez ids (e.g. '123456') or gene symbols (e.g. 'FREB'). Working with such an heterogeneous set of identifiers can be very tedious, particularly due to the lack of any definite one-to-one relationship between them. Consequently many online tools have been developed for mapping and converting ids between the different systems of keys (Alibés et al. 2007; Allen et al. 2011; Bussey et al. 2003; Diehn 2003). We developed a flexible interface to run a pipeline based on Entrez ids and Bioconductor annotation packages (*annotate: Annotation for microarrays*; Gentleman et al. 2004), that converts marker gene lists – and more generally biological/manufacturer identifier lists – between different types of ids. This greatly facilitates the usage of maker lists across multiple microarray platforms or organisms. As starting point, we employed a relatively simple mapping strategy, which only allowed for retrieving either all matching probes or only the first one, including a slightly improved heuristic for Affymetrix probeset identifiers, whose format contains some information on the probeset specificity (e.g. '*_s_at' IDs designate probesets that share common probes among multiple transcripts from different genes). However, we will look in the future at integrating more complex and robust mapping strategies, such as using BLAST alignment based annotations or consensus annotation (Allen et al. 2011), and investigate data-driven approaches, with the objective of optimising marker gene lists for use with a particular dataset (see Section 4).

Example: from Illumina to Affymetrix We illustrate in the following the use of our conversion pipeline. For instance, suppose one has a marker list such as the ones from HaemAtlas (Watkins et al. 2009), whose original identifiers are probe IDs from the Illumina HumanWG6v2 chip, and wants to use it with data generated on an Affymetrix HG U133 Plus 2.0, such as the dataset GSE11058 (Abbas et al. 2009). This dataset contains gene expression measurements for 4 cell lines of immune origin (Jurkat from T cells, Raji and IM-9 from B cells, and THP-1 from monocyte cells), that were hybridised either alone or in mixtures of known proportions. Figure 8 shows how this conversion can be carried out using the *CellMix* package, in the same time illustrating usage of the registry for expression data and marker lists. We show in Figure 16 the same example with verbose output, which provides more details on what happens during the mapping process.

After conversion the marker list is indexed with Affymetrix IDs that match row names in the expression matrix, and may be used to access the markers' expression values, or in combination with some of the visualisation utilities included in the package. For example, Figure 9 shows bar charts of the content of both the original (Illumina IDs) and converted (Affymetrix IDs) lists, in term of number of markers per cell type. The *R* code that generates the plots is also shown[20]. Due to differences in the platform designs, it is expected that some of the probes cannot be uniquely matched, hence the smaller number of probes available for each cell type after conversion.

2.3.3 Assessing marker expressions in pure samples

Given the range of possible technical and biological variation between experiments, marker genes from a given list may show inconsistent or less cell type specific expression patterns on

[20]For clarity purposes, some aesthetic processing is not shown in the code, e.g. setting graphical parameters `las=2, cex.axis=0.7`.

```
## PRELIMINARY: load data from the internal registries load HaemAtlas
## markers
m <- MarkerList("HaemAtlas")
summary(m)

## <object of class MarkerList>
## Types: 8 ['B-CD19', 'Erythroblast', ..., 'T-CD8']
## Mode: character
## Markers: 2069
## IDtype: .Illumina ['ILMN_1793637', 'ILMN_1663575', ..., 'ILMN_1815673']
## Source: illuminaHumanv2.db
## Breakdown:
##            B-CD19      Erythroblast Granulocyte-CD66b     Megakaryocyte
##               247               322               878               279
##     Monocyte-CD14           NK-CD56             T-CD4             T-CD8
##               205                82                51                 5

# load the data using its key NB: require Internet access on first usage)
e <- ExpressionMix("GSE11058")
# probes are not from the same platform
annotation(e)

## [1] "hgu133plus2.db"

##

# convert Illumina to the IDs used in the expression data (HGU133plus2) ->
# this looks for a single match per probe, dropping non-primary affy
# probes
m_affy <- convertIDs(m, e)
summary(m_affy)

## <object of class MarkerList>
## Types: 8 ['B-CD19', 'Erythroblast', ..., 'T-CD8']
## Mode: character
## Markers: 1659
## IDtype: .Affymetrix ['206513_at', '207655_s_at', ..., '221126_at']
## Source: hgu133plus2.db
## Breakdown:
##            B-CD19      Erythroblast Granulocyte-CD66b     Megakaryocyte
##               174               280               668               250
##     Monocyte-CD14           NK-CD56             T-CD4             T-CD8
##               180                62                41                 4
```

Figure 8: *Sample code for converting the HaemAtlas marker list defined by Watkins et al. (2009), from Illumina HumanWG6v2 probe identifiers to Affymetrix HG U133 Plus 2.0 probeset ids. This code also illustrates the usage of the internal registries for expression data and marker lists.*

```
bp <- barplot(m, main = annotation(m))
barplot(m_affy, ylim = bp$ylim, main = annotation(m_affy))
```

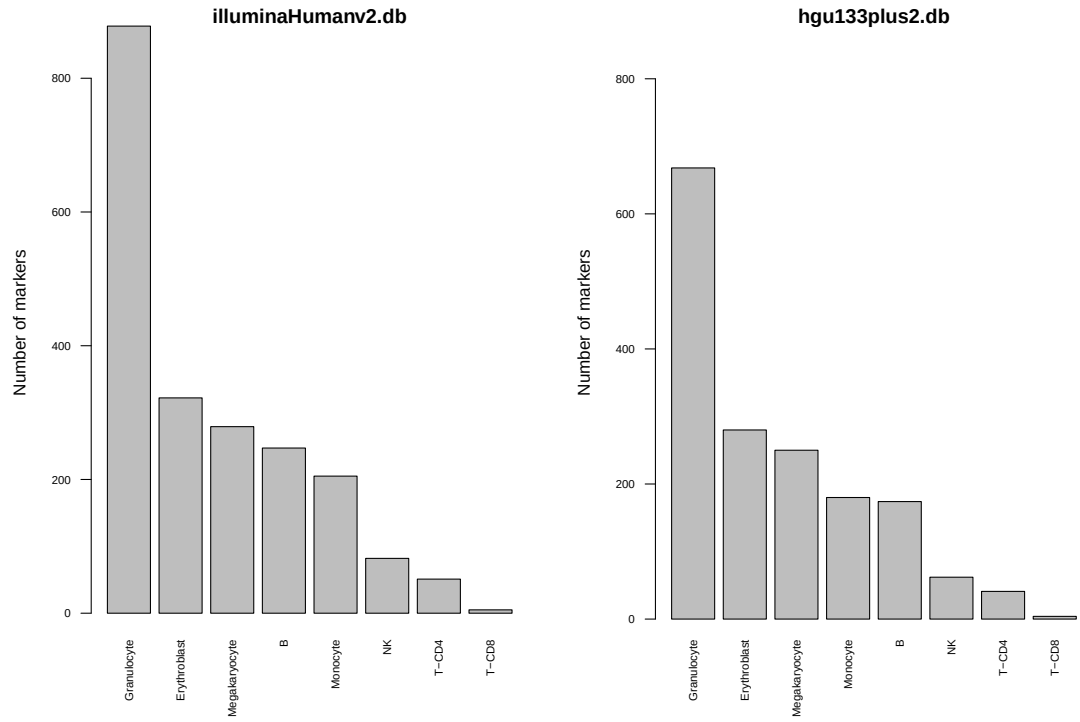


Figure 9: *Number of markers for each cell type in the original HaemAtlas marker list from Watkins et al. (2009) that uses Illumina HumanWG6v2 probe IDs (left panel) and in the list converted to match probesets on Affymetrix HGU 133 Plus 2.0 (right panel).*

an independent dataset. When the considered dataset contains expression values from pure cell types, for which markers are present in the list, one may qualitatively evaluate their consistency with their respective markers, by visualising the expression profiles of these. We illustrate this assessment method using the same dataset and the same marker list as in the previous section, that is, respectively, the GSE11058 dataset from Abbas et al. (2009), in which the pure samples are from transformed cell lines from immune origin, and the list of markers from HaemAtlas (Watkins et al. 2009), which contains markers for a variety of blood cell types (T cells, B cells, Monocytes, etc.). Importantly, these markers were defined completely independently from the cell line dataset. Looking at the markers' expression across these samples, provides insight on how similar the cell lines are to their cell type of origin, and vice-versa.

Figure 10, shows for each cell type in HaemAtlas, the average expression values of the 50 most uniformly highly expressed markers in each cell line. More precisely, each marker is first attributed a measure of uniform expression, computed as the maximum of its minimum expression within each cell line, and then ordered decreasingly according to this value, within its respective cell type in the marker list. The *CellMix* package makes these computations very easy, via the `reorder` function which implements by default this maximin score, and allows for custom scoring schema. The specialised subsetting operator `[]` defined for the marker list data structure, enables the 50 top scoring markers to be extracted separately for each cell type. Besides, because the dataset was retrieved from the internal registry which stored information about which sample is pure, the expression data for the pure samples only are readily extracted using the function `pureSamples`.

The expression patterns in Figure 10 are very different from what would be expected given the cell type of origin of each cell line. Expression values for the complete list of markers, ordered in the same way, show similar patterns (Figure 18). Erythroblast, Granulocyte and Megakaryocyte

markers (yellow and greens) appear to be highly expressed at similar levels by all cell lines, with THP-1 cells – which are from monocytes – expressing Granulocyte markers at a relatively higher level. This is surprising since Erythroblast and Megakaryocyte are precursor cells for red blood cells and platelets respectively, which are completely different lineages from monocytes and lymphocytes. In terms of hematopoiesis, monocytes are closer to Granulocytes than B cells and T cells, but are still very distinct cell types (cf. the hematopoietic tree Figure 14 in Appendix A). In any case, the main concern here comes from the fact that Watkins et al. (2009) originally selected these markers for their ability to distinguish between all these cell types. Consequently, one would expect each cell line to express only the markers from their cell type of origin. This discrepancy could be explained by the fact that the cell lines Raji, Jurkat, THP-1 and IM-9 were originally cloned from cancer immune cells (Epstein et al. 1966; Fahey et al. 1971; Schneider et al. 1977; Tsuchiya et al. 1980), and are therefore not quite equivalent to primary immune cells, especially in terms of gene expression. Plots such as the bar chart in Figure 10 provide a way to visually assess how much they differ in reality, through the prism of known marker genes.

This more generally raises a potential caveat of using marker genes for the analysis of gene expression in – heterogeneous – diseases. Indeed, the disease under study might well affect the expression of the *a priori* chosen marker genes, possibly non uniformly across samples, even within disease groups due to unknown or unannotated disease subtypes. Clearly, including in the sets of markers genes whose expression is not cell type-specific in the particular conditions would jeopardise the estimation and identification of signatures and proportions. It is therefore important to – be able to – assess the discriminating power of a given list of genes in the context of the data to be deconvolve (see Section 4).

As a control step, we plotted in Figure 11 the expression values of markers from the same list in the basis signatures from Abbas et al. (2009). We recall from Section 2.2, that these signatures were defined by Abbas et al. to represent the cell-specific gene expression of 6 immune cell types, in different activation states (e.g. helper T cells (Th), activated helper T cells (Th act)), optimised for performing accurate detailed deconvolution of blood samples. Importantly again, these signatures were defined from microarray gene expression data, completely independently from the marker genes in HaemAtlas. This time, the expression patterns are mainly as expected, with the exception of one Natural killer cell (NK) marker expressed by the cytotoxique T cell signature, and few Granulocyte markers expressed in the monocyte signatures. This increases confidence in both the considered marker gene list and signatures.

2.4 Deconvolution methods

The *CellMix* package provides access to a range of gene expression deconvolution methods, in such a way that they can easily be applied on commonly available data. Combined with the framework developed for benchmark expression data and marker gene lists, this facilitates the comparison and development of methodologies.

Amongst the methods we reviewed in Section 1.1, we included the quadratic programming method from Gong et al. (2011), the *csSAM* least-squares algorithm from Shen-Orr et al. (2010), the Bayesian algorithm *DSection* from Erkkila et al. (2010), the unsupervised NMF method *deconf* from Repsilber et al. (2010), as well as our semi-supervised approach (Gaujoux et al. 2011). Again, all algorithms are stored in an extendible internal registry and convenient interface functions are provided. Table 2 lists all built-in algorithms, along with their respective requirements in terms of input data and iterations (i.e. single or multiple iterations), which is related to the associated computational cost. The first column contains the access key to run the algorithm in *CellMix* package. Columns *Basis*, *Coef* and *Markers* indicate which input data is required for running each algorithm (Required: ✓, Not required: -). The last column indicates how many default iterations the algorithm performs to fit a model. A "-" means that the algorithm is not iterative and generally runs relatively fast; a positive value means that an iterative process is used to estimate the unknown data and may be computationally intensive. For the MCMC algorithm *DSection*, this corresponds to the number of samples performed after the burn-in period, which defaults to 4 times the number of samples. For the NMF algorithms *deconf*, *ssBrunet* and *ssLee*,

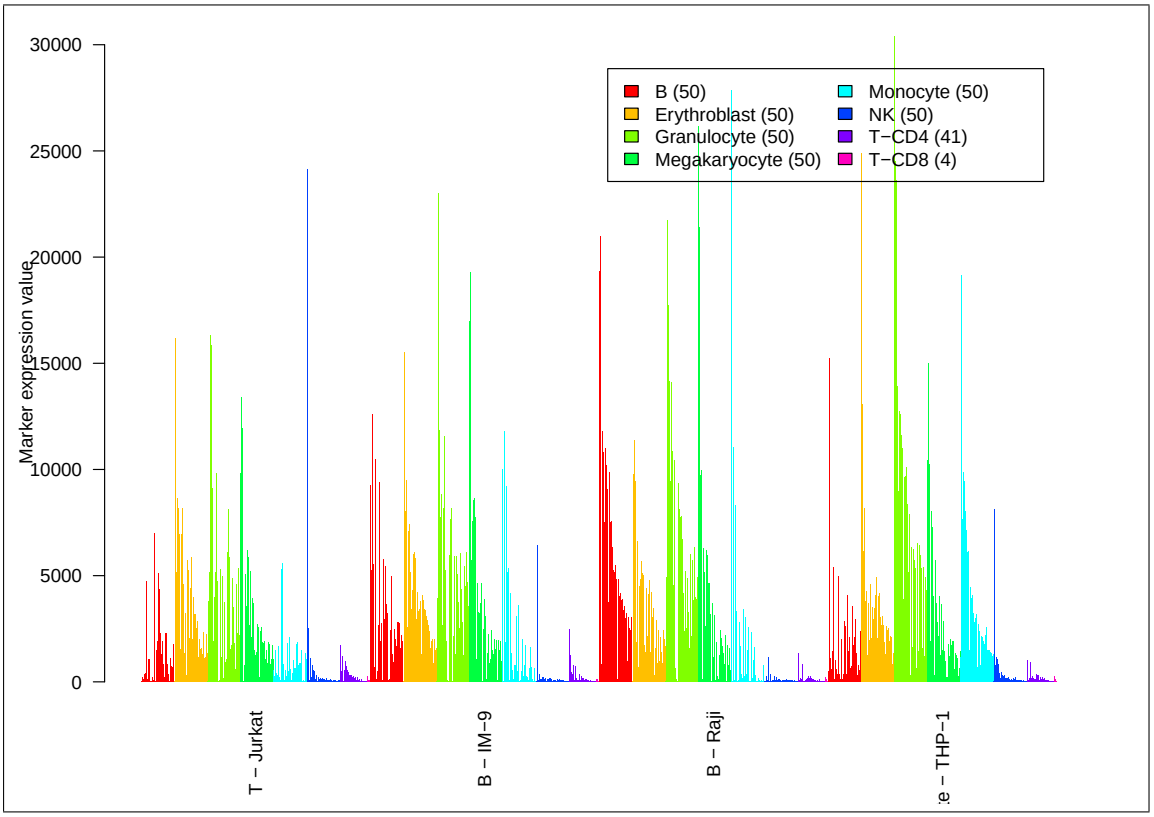


Figure 10: *Expression values of the 50 genes of each cell type in HaemAtlas (Watkins et al. 2009), that are the most uniformly highly expressed in each immune cell line from dataset* `GSE11058`.

this corresponds to the number of block-descent minimising iterations.

2.4.1 Automatic method selection

From a practical point of view, we notice that the choice of an applicable method is essentially determined by the type of data available and the computational resources one is prepared to use. This enables the implementation of a simple procedure for automatically choosing a suitable method, given the input data and the number of possible iterations, and performing gene expression deconvolution in all common situations. We describe here below this procedure, which is implemented within the main interface function `ged` that internally dispatches to the relevant method. Obviously, any deconvolution method may also be explicitly specified, which bypasses the automatic choice procedure.

Known signatures If reliable cell type-specific signatures are known, then estimating the mixture proportions using a standard least-squares approach to linearly regress the global gene expression on the signatures constitute a sensible choice, which has proven to lead to robust results (Abbas et al. 2009; Gong et al. 2011), for a small computational cost (see also Figure 5). The default in this case is to use the quadratic programming approach from Gong et al. (2011), which is based on a more theoretically grounded constrained optimisation technique, which properly handles the nonnegativity and sum-up-to-one constraints on the proportions.

Known proportions If mixture proportions have been measured with good accuracy, then a similar linear regression approach for estimating the signatures may be applied. In this case, partial deconvolution is performed using the *csSAM* algorithm from Shen-Orr et al. (2010), which

```
# convert Illumnina marker ids to match Abbas feature id type
ma <- convertIDs(m, Abbas, method = "all")
# plot markers expression values in Abbas dataset
barplot(ma, Abbas, las = 2, ylab = "")
```

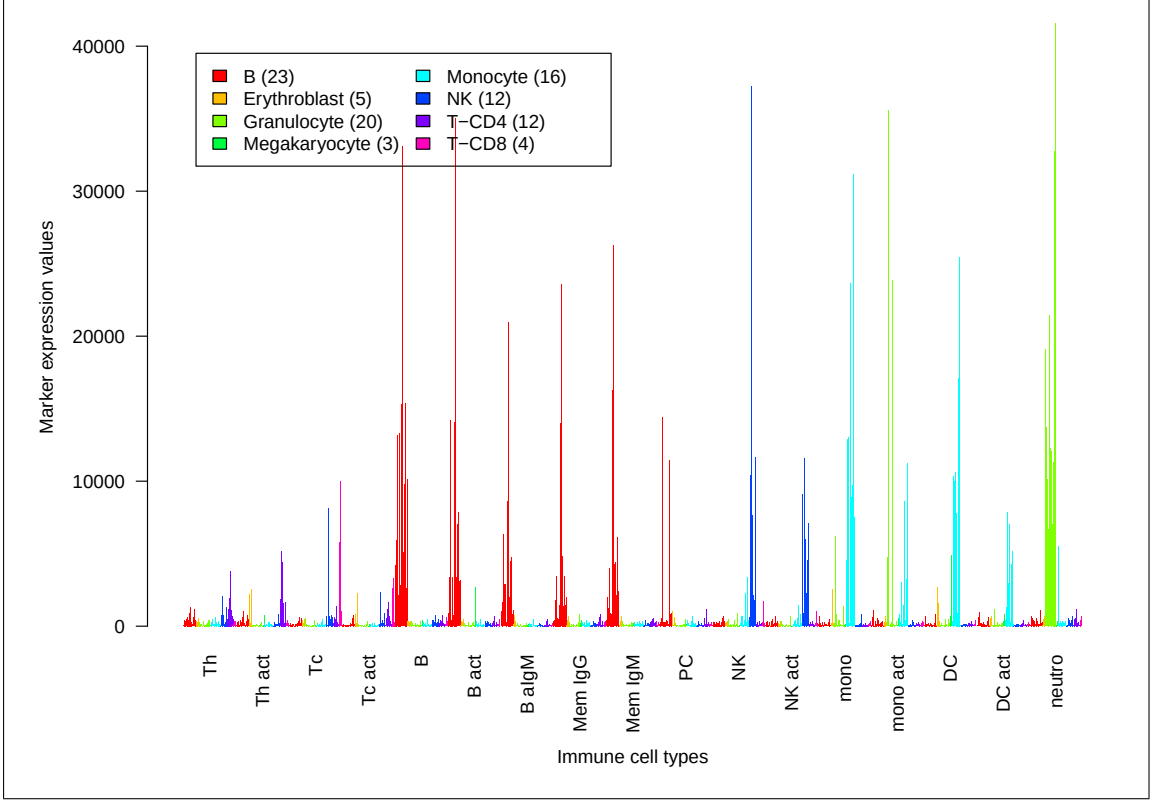


Figure 11: *Expression values of the HaemAtlas marker genes (Watkins et al. 2009) in the basis signature matrix from Abbas et al. (ibid.). The bar chart shows the expression values for the 3529 probesets in the basis matrix that could be mapped by Entrez ID to markers from the list.*

```
basismarkermap(mo[, 1:50], avg, Rowv = NA)
basismarkermap(ma, Abbas, Rowv = NA)
```

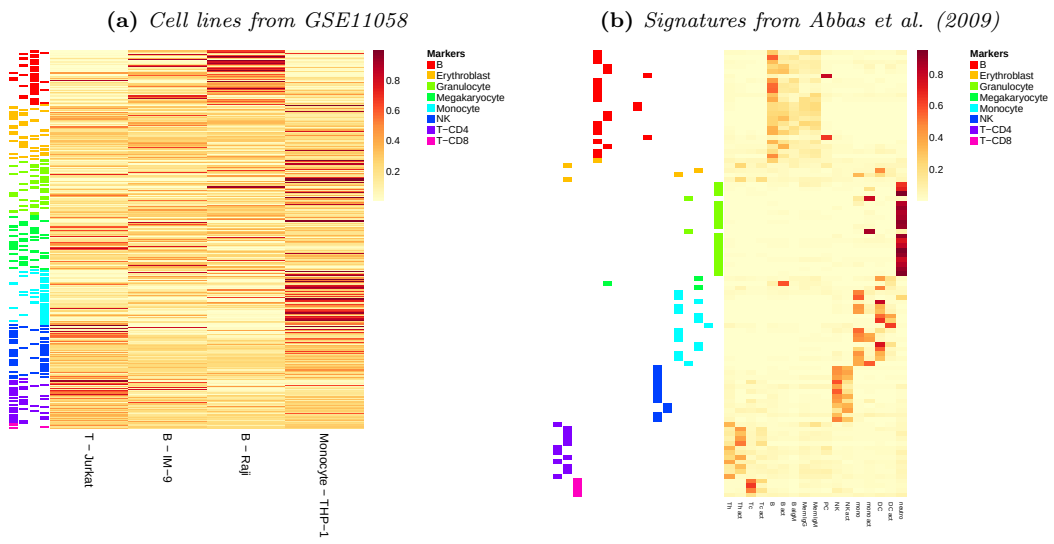


Figure 12: *Heatmap of expression values of the HaemAtlas marker genes in cell lines from* `GSE11058` *and the basis signature matrix from Abbas et al. (ibid.). (a) shows the average expression values for 50 marker genes, as computed in Figure 10. (b) shows the expression values of all markers in each cell type included in the immune signatures.*

in addition is able to compute statistics of differential expression at the level of each cell-type if provided with two groups of samples (e.g. case/control).

Uncertain proportions If mixture proportions are known only to a certain degree of confidence (e.g. measurement errors or rough *a priori* estimation), then the Bayesian *DSection* algorithm (Erkkila et al. 2010) is applied using these prior estimates, to estimate cell-specific signatures, together with adjusted mixture proportions that are more consistent with the observed global expression data. If multiple groups of samples are defined, this algorithm computes signatures and standard error estimates for each group separately, suitable for performing pairwise differential expression analysis. Note that it is technically not possible to decide which of the *csSAM* or *DSection* algorithms should be used, only based on the main input data (a matrix of proportions), whose accuracy is not intrinsically defined. The choice of methods then relies on the number of iterations requested by the user, requiring multiple iterations meaning that the proportions should be considered uncertain and used as priors in a Bayesian framework, rather than as accurate observations in a standard regression model.

Known marker genes If only a list of marker genes is available, without their expression levels in each cell type, then complete deconvolution is performed using our semi-supervised approach (Gaujoux et al. 2011). Although the *deconf* algorithm (Repsilber et al. 2010) is also applicable in this case, we have shown the benefit of incorporating the marker information within the fitting process, as opposed to using it *a posteriori* (Gaujoux et al. 2011). Moreover, as it is currently implemented, the *deconf* algorithm is computationally too demanding on large datasets to be used by default.

Global expression only Given the performances of the *deconf* algorithm in a completely unsupervised setting (ibid.), it is used when no other data than the global expression measurements are

	Description	Basis	Coef	Marker	Iter
lsfit	Partial deconvolution of proportions using standard least-squares (Abbas et al. (2009))	✓	-	-	-
cs-lsfit	Partial deconvolution of cell signatures using standard least-squares	-	✓	-	-
nnls	Partial deconvolution of proportions or signatures using fast combinatorial nonnegative least-squares	✓	-	-	-
cs-nnls	Partial deconvolution of signatures using fast combinatorial nonnegative least-squares	-	✓	-	-
qprog	Estimates proportions from known expression signatures using quadratic programming (Gong et al. (2011))	✓	-	-	-
cs-qprog	Estimates constrained cell-specific signatures from proportions using quadratic programming [experimental]	-	✓	-	-
DSA	Complete deconvolution using Digital Sorting Algorithm (Zhong et al. (2013))	-	-	✓	-
csSAM	Estimates cell/tissue specific signatures from known proportions using SAM (Shen-Orr et al. (2010))	-	✓	-	-
DSection	Estimates proportions from proportions priors using MCMC (Erkkila et al. (2010))	-	✓	-	500
ssKL	Semi-supervised NMF algorithm for KL divergence, using marker genes (Gaujoux et al. (ibid.))	-	-	✓	3000
ssFrobenius	Semi-supervised NMF algorithm for Euclidean distance, using marker genes (Gaujoux et al. (ibid.))	-	-	✓	3000
meanProfile	Compute proportion proxies as mean expression profiles	-	-	✓	-
deconf	Alternate least-square NMF method, using heuristic constraints [fast] (Repsilber et al. (2010))	-	-	-	1000

Table 2: *Algorithms for gene expression deconvolution available in the CellMix package. The first column contains the access key to run the algorithm with the main interface function* `ged`*. Required data are indicated by a ✓. The values in column* Iter *correspond to the default number of iterations performed. A "-" means no intensive iterative process is used.*

available. Alternatively the *ssBrunet* and *ssLee* algorithms may be used in this situation. Since no markers are used in this case, they are equivalent to the standard NMF algorithms from Brunet et al. (2004) and Lee et al. (2001), using the stationarity of the objective function as stopping criterion.

2.4.2 Unified versatile interface

The main interface to run any deconvolution algorithms is implemented in the `ged` function, whose main arguments are (1) the global gene expression data, supporting all commonly used *R* types for matrix-like objects (`matrix`, `data.frame`, `ExpressionSet`); (2) the input data, which can either be a matrix-like object containing known signatures or proportions, or a list of maker genes, or the number of cell/tissue from which signatures and/or proportions must be estimated;

```
# expression data only: deconf
ged(X, 3)
# with markers only: qprog
ged(X, marks)
# with markers only, iterative: ssBrunet
ged(X, marks, maxIter = 1000)
# with known signatures: qprog
ged(X, sig)
# with known proportions: csSAM
ged(X, prop)
# with uncertain proportions, iterative: DSection
ged(X, prop, maxIter = 500)

# force a given algorithm, with extra parameter: ssLee
ged(X, 2, method = "ssLee", markers = marks, ratio = 3)
```

Figure 13: *Sample code for running any deconvolution algorithm. Each line results in a call to a different method, which is determined according to the input data and the number of desired iterations. The last line explicitly specifies a method and extra parameters to form a more complex call.*

(3) the number of iterations to perform; (4) optionally the name of the deconvolution method to use. When no method is specified, it is inferred using the procedure described in the previous section, based on the type and dimensions of the input data (argument (2)) – and in some cases the number of iterations (see above).

Figure 13 shows sample code that deconvolves a given gene expression data object `X`, of dimension $n \times p$, using different kinds of input data. The objects `marks`, `sig` and `prop` are assumed to be respectively a list of marker genes for r cell types, a matrix of expression signatures of dimensions $n \times r$, or a matrix of known proportions of dimension $r \times p$ – with $r < p$. The last line illustrates how one can specify which method should be used and perform more complex calls. In this latter case, signatures and proportions for only the two first cell types in the marker list would be estimated by the method *ssLee* – that would normally not have been called given the kind of input data (default being to use *ssBrunet*). The parameter `ratio` is specific to this method, and controls the expression level threshold of marker genes in their corresponding cell types (e.g. `ratio=2` requires each marker to be expressed by their respective cell type at least three times more than by other cell types).

3 Methods

This section provides relevant implementation details on the internal registries, how the data objects for benchmark datasets are effectively built from their original format, how the marker gene lists were built from the different public databases and microarray studies, and finally how each deconvolution method is implemented.

3.1 Internal registries

The *CellMix* package makes an intensive usage of internal registries to organise, store and manage the different kind of data that are at the core of the package, i.e. benchmark datasets, market gene lists and deconvolution algorithms. For this purpose we use the *registry* package[21] (Meyer 2012) which provides the basic generic infrastructure for registries, with convenient access methods. This package allows to define sets of data fields, with the possibility to control their types,

[21] http://cran.r-project.org/package=registry

validity and access-rights. Its simplicity of use makes it very easy to integrate extensible features with plugin-like capabilities in any *R* package.

3.2 Loading pipeline for benchmark datasets

An internal registry stores for each benchmark dataset the following set of pre-processing functions that compose a pipeline which combines both global expression and cell type specific data into a single data object:

1. Load the expression data, possibly downloading it from its web repository;
2. Filter (optional), e.g. to remove data not related to pure or mixtures samples;
3. Extract/add phenotypic or feature annotation data, e.g. mixture proportions;
4. Create ground truth reference signatures.

The loading step (1) is common to all datasets, and builds upon features from the *GEOquery* package, that downloads files directly from the GEO repository. A transparent caching system reduces the loading time of subsequent access to the data, by storing the original expression data download locally as *.rds* files (a compact binary R format). The optional filtering step (2) enables to subset the original data, to only retain relevant features or samples. For example, the dataset *GSE5350*[22] contains samples from a controlled mixture experiment, as well as from normal/tumour colon samples, which are removed by the filtering step, that in this case keeps only the samples whose source name starts with `'MAQC'` (see Figure 3).

Once the data filtered, meaningful phenotypic information is extracted from different sources, prioritarily focusing on the constituting cell/tissue types and their mixture proportions. For controlled mixture or titration experiments, we extract mixture proportion data directly from the publicly available material (experiment description, data fields, supplementary data). For some of the clinical sample datasets however, CBC or flow cytometry data were not directly available online, or were available only as aggregated results (e.g. mean by case/control group), or with ID mapping issues. We obtained these from the authors, who quickly responded and provided us with the necessary information[23].

The last processing step consists in defining the ground truth reference gene expression signatures for each constituent, which are stored together as an NMF model. For simple experiment designs, that contain expression data from pure samples, these signatures are computed as the mean expression values among each cell type separately. No reference signatures were computed for experiments with more complex designs such as the time course experiments in GSE22886[24] (Abbas et al. 2005) and GSE24223[25] (Grigoryev et al. 2010), but users can relatively easily compute condition-specific signatures, using the relevant phenotypic data made readily available by the annotation extraction step (3).

3.3 Marker gene lists

Marker lists from the IRIS, TiGER and VeryGene databases were extracted from the complete data which can be downloaded from the respective websites. Each one of these database defines marker specificity scores, that are conserved in the internal registry.

Marker lists for Human and Rat tissues from TissueDistributionDB were created using the web interface and the queries shown in Figure 15, which returned a total of 70361 and 22342 entries respectively. Each entry is constituted by a pair (UniGene ID, Tissue type), associated with

[22] For this dataset, we only consider the data from the *GPL570* platform, i.e. Affymetrix Human Genome U133 Plus 2.0 Array, that is accessible separately.

[23] We thank Stephen Popper, Alexander Abbas, Shai Shen-Orr and Alexei Grom for the data they provided.

[24] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22886

[25] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24223

two numeric values (1) The percentage – expression – value[26], which corresponds to the relative proportion of ESTs from the UniGene cluster in the tissue type, normalised across all tissues. (2) the tissue specificity index[27], which indicates how many other tissues also express ESTs from the UniGene cluster. The data for all these entries are available in the package's internal registry, including their respective percentage value and specificity index. By default, marker genes for a given tissue are defined as those whose percentage value is greater than 99%, i.e. those whose cumulated relative expression in all other tissues is at most 1 percent of the relative expression in the assigned tissue. Markers are then ordered by decreasing percentage value and increasing tissue specificity index.

The marker list with access key `Abbas` was extracted from the refined subset of immune genes that compose the basis deconvolution matrix from Abbas et al. (2009). We assigned each probeset to its most expressing cell type, and computed the sparseness of its expression profile across cell types, that is used as a specificity score and defined by Hoyer (2004) as:

$$Sparseness(x) = \frac{\sqrt{n} - \frac{\sum |x_i|}{\sqrt{\sum x_i^2}}}{\sqrt{n} - 1},$$

where in our case x is the expression profile of a given probeset, and n the number of cell types. It is equal to 1 if and only if x contains a single nonzero component, meaning that the gene is expressed by a single cell type, and is equal to 0 if and only if all components of `x` are equal, meaning that the gene is uniformly equally expressed by all cell types.

The marker list from HaemAtlas (Watkins et al. 2009) was extracted from the Supplementary Table 5 of the related paper[28]. Curiously, although they are supposed to be relative to Illumina HumanWG-6 version 2 Expression BeadChip arrays, none of the probe ids found in the table match the Illumina probe ids used in the corresponding `illuminaHumanv2.db` annotation package[29] (*illuminaHumanv2.db: Illumina HumanWG6v2 annotation data (chip illuminaHumanv2)*) from Bioconductor. However, since these data provide the nucleotide sequence of each probe, we could uniquely match all of them by nuID (Du et al. 2007) to the probe IDs from the annotation package. The latter were eventually used as identifiers in the marker gene list. No specificity score is available for this marker list.

The marker list from Palmer et al. (2006) was extracted from the supplementary file associated with the paper[30]. The internal registry contains for each marker two numeric values that may be used as specificity score: (1) the correlation with the theoretical abundance of the assigned cell type; (2) expression fold-change from the differential analysis between cell types carried out by Palmer et al. No default filtering is applied.

3.4 Deconvolution methods

All algorithms are implemented within the framework defined by the *NMF* package (Gaujoux et al. 2010), some purely in *R*, others using optimised *C++* routines, with the exception of *DSection*, whose original *Matlab* code[31] was slightly adapted to make it compatible with *Octave* and is run through the *RcppOctave* package[32] (Gaujoux 2013). Both the *lsfit* and *csSAM* algorithms are implemented using the *R*-core function `lsfit`, which can fit standard least-squares with multiple left-hand sides. The *qprog* algorithm implements the approach from Gong et al. (2011) using the function `lsei` from the *limSolve* package[33] (Soetaert et al. 2009; Van den Meersche et al. 2009), which solves least-squares problems with equality and inequality constraints as a quadratic programming problem, using the algorithm from Haskell et al. (1981). Each global expression

[26] http://genome.dkfz-heidelberg.de/menu/tissue_db/faq.html#percent
[27] http://genome.dkfz-heidelberg.de/menu/tissue_db/faq.html#speci
[28] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2680378/bin/blood-2008-06-162958_TableS5.xls
[29] http://www.bioconductor.org/packages/release/data/annotation/html/illuminaHumanv2.db.html
[30] http://www.biomedcentral.com/content/supplementary/1471-2164-7-115-S2.xls
[31] http://www.cs.tut.fi/~erkkila2/software/dsection/DSection.m
[32] http://cran.r-project.org/package=RcppOctave
[33] http://cran.r-project.org/package=limSolve

profile is solved separately. Our implementation optionally allows for incomplete signatures, i.e. for some constituents to be absent from the signatures, by allowing the sum of the estimated proportions to be lower than 1 – rather than equal to 1 in the standard case. The *ssKL* and *ssFrobenius* algorithms are modifications of NMF algorithms from Brunet et al. (2004) and Lee et al. (2001) respectively, as described in (Gaujoux et al. 2011). The *deconf* algorithm uses the original implementation found in the *deconf* package[34] from Repsilber et al. (2010).

4 Discussion

The use case examples shown in this vignette demonstrate that the *CellMix* package provides a powerful flexible computational framework for gene expression deconvolution applications, whether one wants to simply estimate proportions and/or cell specific signatures from particular gene expression data, apply a specific method, or develop new methodologies. The internal registries offer an easy and managed access to several marker gene lists and real benchmark datasets that contain reference ground truth proportions and/or signatures, which greatly facilitates method comparisons and data integration from multiple sources. The unified interface for running deconvolution algorithms, combined with the automatic method selection, simplifies and improves the user experience, especially for non-advanced users. Moreover, the data structure we developed for storing marker gene lists, comes with a set of utility functions that help in the creation and manipulation of such data. In this section we discuss some possible improvements and challenges that remain to be addressed.

4.0.1 Interface & data

More advanced capabilities for comparing, combining and visualising the content of marker gene lists would be useful. Working towards the easy integration of multiple lists would help building more complete and robust lists. With respect to the implementation of these functionalities, the functions already available for the marker list data structures should provide a good base to build on. Existing strategies for handling multiple lists of genes produced by differential expression analysis should also be considered (Lottaz et al. 2006; Yang et al. 2007).

It is also desirable that more datasets and marker lists are included in the package, as they become available. For immune cells in particular, primarily focus would be on integrating data from some of the specialised databases the listed in Gardy et al. (2009) – of which we became aware only recently. A potential fruitful enhancement would be to enable the direct interaction with those databases that provide a programmatic interface, in a similar way the *GEOquery* package, *ArrayExpress* package or *biomaRt* package[35] (Durinck et al. 2009) do with their respective online databases.

4.0.2 Deconvolution basis matrices

Given the good results achieved by the partial deconvolution approach from Abbas et al. (2009), it would be interesting to build deconvolution basis matrices for other types of cells, tissues and organisms. This could be implemented as a generic pipeline run on extractions of the constantly growing compendium of public datasets available, and assessed based on studies where cell/tissue proportions are known. More generally, investigating methodologies to select and optimise the set of genes/probesets that compose such deconvolution basis matrices is of prime interest, so as to further increase the estimation accuracy.

With respect to this, the approach based on the condition number of the deconvolution basis matrix used by Abbas et al. deserves some attention. In general optimisation theory, the condition number is a notion associated with any optimisation problem, that measures the sensitivity of the solution(s) with respect to the data, and is related to numerical stability of computed solutions and

[34] http://www.biomedcentral.com/content/supplementary/1471-2105-11-27-s1.zip
[35] http://www.bioconductor.org/packages/release/bioc/html/biomaRt.html

backward error estimation (Bertsekas 1999; Higham 1996). Essentially, the smaller the condition number, the more the solution computed by a given algorithm is likely to be robust to variations in the data, although strictly speaking the sensitivity to variations is intrinsically dependent on the algebraic and optimisation techniques used by the algorithm. In particular, the condition number of least-square problems is proportional to the condition number of the coefficient matrix (Grcar 2003), i.e. the deconvolution basis matrix in the case of partial deconvolution. This explains the correlation between accuracy and condition number observed by Abbas et al. Indeed, they built the different basis matrices using expression data from pure samples. Due to measurement errors and cell interactions (Cobb et al. 2005; Palmer et al. 2006; Patocs et al. 2007), global gene expression profiles from mixed samples deviate from the theoretical expression profiles, which consist of the exact mixture of the pure profiles (i.e. the columns of the basis matrix). Being associated with better conditioned optimisation problems, matrices with smaller condition numbers are therefore expected to estimate proportions that are closer to the actual proportions.

4.0.3 Data-driven optimisation of marker gene lists

When pure samples are available in an experiment, the consistency of a given list of marker genes may be assessed by comparing the markers' expression profiles with the expectation of block expression patterns. Bar charts such as in Figure 10 or heatmaps such as in Figure 12 help pinpoint inconsistencies. Quantitative measures would however be useful in this context, so that the consistency of each marker can be computed, and used to automatically select good markers. Differential expression statistics used to define markers can typically be used as consistency measures, although the usually small number of pure samples from each cell type might suggest the use of heuristic scoring schemas.

However, a use case where marker gene lists are most needed is when the considered experiment contains no data from pure samples. One would therefore want to be able to refine a given marker gene list, so that it is consistent with the expression data of the mixed samples. Note that this would in fact be applicable and relevant even when pure samples are available.

Very recently, Schneider et al. (2011) proposed a statistical test based on Kendall's W coefficient of concordance to identify concordant redundant probesets in expression microarray data. Their approach, named Statistical Consolidation of Redundant Expression Measures (SCOREM), is based on the observation that groups of probesets annotated as representing the same gene may have discordant expression profiles, due to cross-hybridisation, misannotation or alternate-splicing. Traditional approaches use either one or all of these multiple probesets-per-gene to compute a single representative expression value for each gene, e.g. based on their individual variations or their connectivity in co-expression networks (Miller et al. 2011). The actual expression data from many probesets are therefore not fully used, which lowers the power of detection of biologically meaningful expression patterns. By grouping these probesets into sub-groups whose expression profiles are highly consistent, SCOREM provides clearer biologically relevant signals, enhancing the results of downstream analysis. One interesting aspect of this approach is that it is dataset specific, considering each probeset in the context of its expression profiles in the given experimental conditions, rather than assuming globally applicable expression patterns, from which are derived generic annotation files.

Drawing a parallel between groups of probesets used to represent the same gene, and groups of marker genes used to represent the same cell or tissue type, the SCOREM approach could be used to extract from a marker gene list the most suitable sub-groups of candidate markers for the deconvolution of a particular dataset. Indeed, regardless of the phenotypic characteristics of mixed samples present in the dataset (e.g. case/controls, tumour/normal), good marker genes in each cell/tissue type should have concordant expression profiles, being presumably well correlated with the cell/tissue relative proportion. Given a list of marker genes and expression data from mixed samples only (e.g. clinical blood samples), applying the SCOREM algorithm to each set of markers present in the dataset, and choosing the most concordant or biggest sub-group would create a (sub)list of markers optimised for that specific dataset.

This strategy could prove to be very beneficial to deconvolution methods that make use of

markers (Gaujoux et al. 2011; Repsilber et al. 2010), while other methods may achieve greater accuracy if applied on the global expression data of the subset of highly consistent markers only. Another interesting potential advantage is that this could solve altogether the issue of non-unique probe mapping when converting marker lists across platforms or organisms. One would simply perform this data-driven filtering directly on the list of all mapped probes (e.g. by Entrez IDs) within their respective cell-types. This is similar in essence to the approach used by Miller et al. (2011) for meta-analysis and cell proportion estimation, with the difference that one would search for multiple rather than single representative probesets. The approach may also be used to assess the consistency of estimated cell type-specific signatures. Future work will include evaluating and validating this approach, and eventually integrating it with the other tools provided by the *CellMix* package[36].

5 Conclusion

We developed the *CellMix* package, a general framework for gene expression deconvolution, which provides an easy access to benchmark data, marker gene lists and a variety of deconvolution methods. This package dramatically facilitates deconvolution analysis, by addressing the common practical issues that arise when performing such analysis. Overall, it allows for smoother and more generic analysis pipelines – as well as a friendlier user experience.

The *CellMix* package is currently in alpha version and available on our CRAN-like repository `http://web.cbio.uct.ac.za/~renaud/CRAN`, from which it can be installed as a source package using the standard procedure. We will eventually submit the package to Bioconductor[37] (Gentleman et al. 2004), where it should reach and benefit the wider bioinformatics research community.

[36] A first experimental implementation of the SCOREM approach applied to marker gene filtering is already available in `CellMix`

[37] `http://www.bioconductor.org`

References

[1] Alexander R Abbas, Kristen Wolslegel, Dhaya Seshasayee, Zora Modrusan, and Hilary F Clark. "Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus." In: *PloS one* 4.7 (Jan. 2009), e6098. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0006098. URL: http://www.ncbi.nlm.nih.gov/pubmed/19568420 (see pp. 3–8, 10, 12, 13, 15, 16, 18, 19, 21–26, 29–31, 42).

[2] Alexander R Abbas, D Baldwin, Y Ma, W Ouyang, A Gurney, F Martin, S Fong, M van Lookeren Campagne, P Godowski, P M Williams, a C Chan, and H F Clark. "Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data." In: *Genes and immunity* 6.4 (June 2005), pp. 319–31. ISSN: 1466-4879. DOI: 10.1038/sj.gene.6364173. URL: http://www.ncbi.nlm.nih.gov/pubmed/15789058 (see pp. 10, 18, 28).

[3] Andreu Alibés, Patricio Yankilevich, Andrés Cañada, and Ramón Díaz-Uriarte. "IDconverter and IDClight: conversion and annotation of gene and protein IDs." In: *BMC bioinformatics* 8 (2007), p. 9. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-9. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1779800\&tool=pmcentrez\&rendertype=abstract (see p. 19).

[4] Jeffrey D Allen, Siling Wang, Min Chen, Luc Girard, John D Minna, Yang Xie, and Guanghua Xiao. "Probe mapping across multiple microarray platforms." In: *Briefings in bioinformatics* (2011). ISSN: 1477-4054. DOI: 10.1093/bib/bbr076. URL: http://www.ncbi.nlm.nih.gov/pubmed/22199380 (see p. 19).

[5] U Alon, N Barkai, D a Notterman, K Gish, S Ybarra, D Mack, and a J Levine. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays." In: *Proceedings of the National Academy of Sciences of the United States of America* 96.12 (1999), pp. 6745–50. ISSN: 0027-8424. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=21986&tool=pmcentrez&rendertype=abstract (see p. 6).

[6] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly a Marshall, Katherine H Phillippy, Patti M Sherman, Rolf N Muertter, Michelle Holko, Oluwabukunmi Ayanbule, Andrey Yefanov, and Alexandra Soboleva. "NCBI GEO: archive for functional genomics data sets–10 years on." In: *Nucleic acids research* 39.November 2010 (2010), pp. 1005–1010. ISSN: 1362-4962. DOI: 10.1093/nar/gkq1184. URL: http://www.ncbi.nlm.nih.gov/pubmed/21097893 (see p. 7).

[7] D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999. ISBN: 1-886529-00-0. URL: http://www.athenasc.com/nonlinbook.html (see p. 31).

[8] M S Boguski, T M Lowe, and C M Tolstoshev. "dbEST–database for "expressed sequence tags"." In: *Nature genetics* 4.4 (1993), pp. 332–3. ISSN: 1061-4036. DOI: 10.1038/ng0893-332. URL: http://www.ncbi.nlm.nih.gov/pubmed/8401577 (see p. 11).

[9] Christopher R Bolen, Mohamed Uduman, and Steven H Kleinstein. "Cell Subset Prediction for Blood Genomic Studies". In: *BMC Bioinformatics* 12.1 (2011), p. 258. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-258. URL: http://www.biomedcentral.com/1471-2105/12/258 (see pp. 6, 7, 11, 12).

[10] B M Bolstad, R a Irizarry, M Astrand, and T P Speed. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." In: *Bioinformatics (Oxford, England)* 19.2 (2003), pp. 185–93. ISSN: 1367-4803. URL: http://www.ncbi.nlm.nih.gov/pubmed/12538238 (see p. 13).

[11] A Brazma et al. "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." In: *Nature genetics* 29.4 (2001), pp. 365–71. ISSN: 1061-4036. DOI: 10.1038/ng1201-365. URL: http://dx.doi.org/10.1038/ng1201-365 (see p. 7).

[12] L Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32. URL: http://www.springerlink.com/index/U0P06167N6173512.pdf (see p. 6).

[13] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. "Metagenes and molecular pattern discovery using matrix factorization." In: *Proceedings of the National Academy of Sciences of the United States of America* 101.12 (2004), pp. 4164–9. ISSN: 0027-8424. DOI: 10.1073/pnas.0308531101. URL: http://www.ncbi.nlm.nih.gov/pubmed/15016911 (see pp. 26, 30).

[14] Kimberly J Bussey, David Kane, Margot Sunshine, Sudar Narasimhan, Satoshi Nishizuka, William C Reinhold, Barry Zeeberg, Weinstein Ajay, and John N Weinstein. "MatchMiner: a tool for batch navigation among gene and gene product identifiers." In: *Genome biology* 4.4 (2003), R27. ISSN: 1465-6914. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=154578\&tool=pmcentrez\&rendertype=abstract (see p. 19).

[15] Scott L Carter, Aron C Eklund, Brigham H Mecham, Isaac S Kohane, and Zoltan Szallasi. "Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements." In: *BMC bioinformatics* 6 (2005), p. 107. ISSN: 1471-2105. DOI: 10.1186/1471-2105-6-107. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1127107\&tool=pmcentrez\&rendertype=abstract (see p. 17).

[16] Jennifer Clarke, Pearl Seo, and Bertrand Clarke. "Statistical expression deconvolution from mixed tissue samples." In: *Bioinformatics (Oxford, England)* 26.8 (2010), pp. 1043–9. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq097. URL: http://www.ncbi.nlm.nih.gov/pubmed/20202973 (see pp. 3–5).

[17] J Perren Cobb et al. "Application of genome-wide expression analysis to human health and disease." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.13 (2005), pp. 4801–6. ISSN: 0027-8424. DOI: 10.1073/pnas.0409768102. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=555033\&tool=pmcentrez\&rendertype=abstract (see p. 31).

[18] Sean Davis and Paul Meltzer. "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor". In: *Bioinformatics* 14 (2007), pp. 1846–1847 (see p. 7).

[19] M. Diehn. "SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data". In: *Nucleic Acids Research* 31.1 (2003), pp. 219–223. ISSN: 13624962. DOI: 10.1093/nar/gkg014. URL: http://www.nar.oupjournals.org/cgi/doi/10.1093/nar/gkg014 (see p. 19).

[20] Sorin Draghici, Purvesh Khatri, Aron C Eklund, and Zoltan Szallasi. "Reliability and reproducibility issues in DNA microarray measurements." In: *Trends in genetics : TIG* 22.2 (2006), pp. 101–9. ISSN: 0168-9525. DOI: 10.1016/j.tig.2005.12.005. URL: http://www.ncbi.nlm.nih.gov/pubmed/16380191 (see p. 17).

[21] Pan Du, Warren a Kibbe, and Simon M Lin. "nuID: a universal naming scheme of oligonucleotides for illumina, affymetrix, and other microarrays." In: *Biology direct* 2 (2007), p. 16. ISSN: 1745-6150. DOI: 10.1186/1745-6150-2-16. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1891274\&tool=pmcentrez\&rendertype=abstract (see p. 29).

[22] Mark Dunning, Andy Lynch, and Matthew Eldridge. *illuminaHumanv2.db: Illumina HumanWG6v2 annotation data (chip illuminaHumanv2)*. R package version 1.16.0 (see p. 29).

[23] Steffen Durinck, Paul T. Spellman, Ewan Birney, and Wolfgang Huber. "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt". In: *Nature Protocols* 4 (2009), pp. 1184–1191 (see p. 30).

[24] M A Epstein, B G Achong, Y M Barr, B Zajac, G Henle, and W Henle. "Morphological and virological investigations on cultured Burkitt tumor lymphoblasts (strain Raji)." In: *Journal of the National Cancer Institute* 37.4 (1966), pp. 547–59. ISSN: 0027-8874. URL: http://www.ncbi.nlm.nih.gov/pubmed/4288580 (see p. 22).

[25] Timo Erkkila, Saara Lehmusvaara, Pekka Ruusuvuori, Tapio Visakorpi, Ilya Shmulevich, and Harri Lahdesmaki. "Probabilistic analysis of gene expression measurements from heterogeneous tissues." In: *Bioinformatics (Oxford, England)* 26.20 (2010), pp. 2571–7. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq406. URL: http://www.ncbi.nlm.nih.gov/pubmed/20631160http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2951082\&tool=pmcentrez\&rendertype=abstract (see pp. 3–5, 7, 8, 22, 25, 26).

[26] J L Fahey, D N Buell, and H C Sox. "Proliferation and differentiation of lymphoid cells: studies with human lymphoid cell lines and immunoglobulin synthesis." In: *Annals of the New York Academy of Sciences* 190 (1971), pp. 221–34. ISSN: 0077-8923. URL: http://www.ncbi.nlm.nih.gov/pubmed/5290016 (see p. 22).

[27] Rita Ferreira, Kinuko Ohneda, Masayuki Yamamoto, and Sjaak Philipsen. "GATA1 function, a paradigm for transcription factors in hematopoiesis". In: *Molecular and cellular biology* 25.4 (2005), p. 1215. DOI: 10.1128/MCB.25.4.1215. URL: http://mcb.asm.org/cgi/content/abstract/25/4/1215 (see p. 40).

[28] Jennifer L Gardy, David J Lynn, Fiona S L Brinkman, and Robert E W Hancock. "Enabling a systems biology approach to immunology: focus on innate immunity." In: *Trends in immunology* 30.6 (2009), pp. 249–62. ISSN: 1471-4981. DOI: 10.1016/j.it.2009.03.009. URL: http://www.ncbi.nlm.nih.gov/pubmed/19428301 (see p. 30).

[29] Renaud Gaujoux. *RcppOctave: Seamless Interface to Octave – and Matlab*. R package version 0.9. 2013. URL: http://CRAN.R-project.org/package=RcppOctave (see p. 29).

[30] Renaud Gaujoux and Cathal Seoighe. "A flexible R package for nonnegative matrix factorization." In: *BMC bioinformatics* 11 (2010), p. 367. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-367. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2912887/ (see pp. 12, 29).

[31] Renaud Gaujoux and Cathal Seoighe. "CellMix: A Comprehensive Framework for Gene Expression Deconvolution". In: *submitted* (2012) (see p. 1).

[32] Renaud Gaujoux and Cathal Seoighe. "Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study." In: *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* (2011). ISSN: 1567-7257. DOI: 10.1016/j.meegid.2011.08.014. URL: http://www.ncbi.nlm.nih.gov/pubmed/21930246 (see pp. 3, 6, 7, 11, 12, 22, 25, 26, 30, 32).

[33] Pascal Gellert, Katharina Jenniches, Thomas Braun, and Shizuka Uchida. "C-It: a knowledge database for tissue-enriched genes." In: *Bioinformatics (Oxford, England)* 26.18 (2010), pp. 2328–33. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq417. URL: http://www.ncbi.nlm.nih.gov/pubmed/20628071 (see p. 11).

[34] R. Gentleman. *annotate: Annotation for microarrays*. R package version 1.36.0 (see p. 19).

[35] Robert C Gentleman et al. "Bioconductor: open software development for computational biology and bioinformatics." In: *Genome biology* 5.10 (2004), R80. ISSN: 1465-6914. DOI: 10.1186/gb-2004-5-10-r80. URL: http://www.ncbi.nlm.nih.gov/pubmed/15461798 (see pp. 12, 19, 32).

[36] Ting Gong, Nicole Hartmann, Isaac S Kohane, Volker Brinkmann, Frank Staedtler, Martin Letzkus, Sandrine Bongiovanni, and Joseph D Szustakowski. "Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples." In: *PloS one* 6.11 (Jan. 2011), e27156. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0027156. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3217948\&tool=pmcentrez\&rendertype=abstract (see pp. 3, 5, 7, 8, 10, 12, 13, 18, 22, 23, 26, 29).

[37] JF Grcar. "Optimal sensitivity analysis of linear least squares". In: *Lawrence Berkeley National Laboratory, Report LBNL-* 99 (2003). URL: http://scholar.google.com/scholar?hl=en\&btnG=Search\&q=intitle:Optimal+Sensitivity+Analysis+of+Linear+Least+Squares\#0 (see p. 31).

[38] Marion Gremse, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. "The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources." In: *Nucleic acids research* 39.Database issue (2011), pp. D507–13. ISSN: 1362-4962. DOI: 10.1093/nar/gkq968. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013802\&tool=pmcentrez\&rendertype=abstract (see p. 11).

[39] Yevgeniy a Grigoryev et al. "Deconvoluting post-transplant immunity: cell subset-specific mapping reveals pathways for activation and expansion of memory T, monocytes and B cells." In: *PloS one* 5.10 (2010), e13358. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0013358. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2954794\&tool=pmcentrez\&rendertype=abstract (see pp. 7, 18, 28).

[40] K.H. Haskell and R.J. Hanson. "An algorithm for linear least squares problems with equality and nonnegativity constraints". In: *Mathematical Programming* 21.1 (1981), pp. 98–118. URL: http://www.springerlink.com/index/U677717422346122.pdf (see p. 29).

[41] N.J. Higham. *Accuracy and stability of numerical algorithms*. 48. Siam, 1996. URL: http://books.google.com/books?hl=en\&lr=\&id=EbBWmmPcjPIC\&oi=fnd\&pg=PR13\&dq=Accuracy+and+stability+of+numerical+algorithms\&ots=_yDHlSI6-t\&sig=G2w4Qr6z2f2pg7Tp1c9R9ABWzEg (see p. 31).

[42] PO Hoyer. "Non-negative matrix factorization with sparseness constraints". In: *The Journal of Machine Learning Research* 5 (2004), pp. 1457–1469. URL: http://portal.acm.org/citation.cfm?id=1044709 (see pp. 16, 29).

[43] Marc Jacobsen, D Repsilber, A Gutschmidt, A Neher, K Feldmann, HJ Mollenkopf, SHE Kaufmann, and A Ziegler. "Deconfounding microarray analysis". In: *Methods of information in medicine* 45.5 (2006), pp. 557–563. URL: http://www.schattauer.de/de/magazine/uebersicht/zeitschriften-a-z/methods/contents/archive/issue/680/manuscript/7198.html (see p. 5).

[44] Charles A Janeway, Paul Travers, Mark Walport, and Mark J Shlomchik. *Immunobiology : the immune system in health and disease*. Ed. by Charles A Jr Janeway, Paul Travers, Mark Walport, and Mark J Shlomchik. 5th. Vol. 6. Garland Science, 2001. Chap. Chapter 1, pp. 461–488. ISBN: 0815341016. URL: http://www.ncbi.nlm.nih.gov/books/NBK27092/\#A39 (see p. 40).

[45] Audrey Kauffmann, Tim F. Rayner, Helen Parkinson, Misha Kapushesky, Margus Lukk, Alvis Brazma, and Wolfgang Huber. "Importing ArrayExpress datasets into R/Bioconductor". In: *Bioinformatics* 25.16 (2009), pp. 2092–4 (see p. 7).

[46] Sunitha Kogenaru, Coral Val, Agnes Hotz-Wagenblatt, and Karl-Heinz Glatting. "TissueDistributionDBs: a repository of organism-specific tissue-distribution profiles". In: *Theoretical Chemistry Accounts* 125.3-6 (2009), pp. 651–658. ISSN: 1432-881X. DOI: 10.1007/s00214-009-0670-5. URL: http://www.springerlink.com/index/10.1007/s00214-009-0670-5http://www.springerlink.com/index/V03177J8214R3341.pdf (see pp. 11, 18).

[47] Alexandre Kuhn, Doris Thu, Henry J Waldvogel, Richard L M Faull, and Ruth Luthi-Carter. "Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain." In: *Nature methods* 8.11 (2011), pp. 945–7. ISSN: 1548-7105. DOI: 10.1038/nmeth.1710. URL: http://www.ncbi.nlm.nih.gov/pubmed/21983921 (see pp. 3, 6, 7, 11).

[48] Harri Lähdesmäki, Llya Shmulevich, Valerie Dunmire, Olli Yli-Harja, and Wei Zhang. "In silico microdissection of microarray data from heterogeneous cell populations." In: *BMC bioinformatics* 6 (2005), p. 54. ISSN: 1471-2105. DOI: 10.1186/1471-2105-6-54. URL: http://www.ncbi.nlm.nih.gov/pubmed/15766384 (see pp. 3, 5, 6).

[49] Peter Langfelder and Steve Horvath. "WGCNA : an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 13 (2008). DOI: 10.1186/1471-2105-9-559 (see p. 12).

[50] D D Lee and HS Seung. "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems* (2001). URL: http://scholar.google.com/scholar?q=intitle:Algorithms+for+non-negative+matrix+factorization\#0 (see pp. 26, 30).

[51] Shuang Liang, Yizheng Li, Xiaobing Be, Steve Howes, and Wei Liu. "Detecting and profiling tissue-selective genes." In: *Physiological genomics* 26.2 (2006), pp. 158–62. ISSN: 1531-2267. DOI: 10.1152/physiolgenomics.00313.2005. URL: http://www.ncbi.nlm.nih.gov/pubmed/16684803 (see p. 11).

[52] Weixiang Liu, Kehong Yuan, and Datian Ye. "Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis." In: *Journal of biomedical informatics* 41.4 (2008), pp. 602–6. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2007.12.003. URL: http://www.ncbi.nlm.nih.gov/pubmed/18234564 (see p. 11).

[53] Xiong Liu, Xueping Yu, Donald J Zack, Heng Zhu, and Jiang Qian. "TiGER: a database for tissue-specific gene expression and regulation." In: *BMC bioinformatics* 9 (2008), p. 271. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-271. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2438328\&tool=pmcentrez\&rendertype=abstract (see p. 18).

[54] Claudio Lottaz, Xinan Yang, Stefanie Scheid, and Rainer Spang. "OrderedList–a bioconductor package for detecting similarity in ordered gene lists." In: *Bioinformatics (Oxford, England)* 22.18 (2006), pp. 2315–6. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btl385. URL: http://www.ncbi.nlm.nih.gov/pubmed/16844712 (see p. 30).

[55] Peng Lu, Aleksey Nakorchevskiy, and Edward M Marcotte. "Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations." In: *Proceedings of the National Academy of Sciences of the United States of America* 100.18 (2003), pp. 10370–5. ISSN: 0027-8424. DOI: 10.1073/pnas.1832361100. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=193568\&tool=pmcentrez\&rendertype=abstract (see pp. 3, 5, 8).

[56] Brigham H Mecham, Gregory T Klus, Jeffrey Strovel, Meena Augustus, David Byrne, Peter Bozso, Daniel Z Wetmore, Thomas J Mariani, Isaac S Kohane, and Zoltan Szallasi. "Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements." In: *Nucleic acids research* 32.9 (2004), e74. ISSN: 1362-4962. DOI: 10.1093/nar/gnh071. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=419626\&tool=pmcentrez\&rendertype=abstract (see p. 17).

[57] David Meyer. *registry: Registry infrastructure*. R package version 0.2. 2012. URL: http://CRAN.R-project.org/package=registry (see p. 27).

[58] Jeremy a Miller, Chaochao Cai, Peter Langfelder, Daniel H Geschwind, Sunil M Kurian, Daniel R Salomon, and Steve Horvath. "Strategies for aggregating gene expression data: The collapseRows R function". In: *BMC Bioinformatics* 12.1 (2011), p. 322. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-322. URL: http://www.biomedcentral.com/1471-2105/12/322 (see pp. 6, 7, 12, 31, 32).

[59] Chana Palmer, Maximilian Diehn, Ash a Alizadeh, and Patrick O Brown. "Cell-type specific gene expression profiles of leukocytes in human peripheral blood." In: *BMC genomics* 7 (2006), p. 115. ISSN: 1471-2164. DOI: 10.1186/1471-2164-7-115. URL: http://www.ncbi.nlm.nih.gov/pubmed/16704732 (see pp. 16, 18, 29, 31).

[60] Helen Parkinson et al. "ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression." In: *Nucleic acids research* 37.Database issue (2009), pp. D868–72. ISSN: 1362-4962. DOI: 10.1093/nar/gkn889. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686529\&tool=pmcentrez\&rendertype=abstract (see p. 7).

[61] A. Patocs, L. Zhang, Y. Xu, F. Weber, T. Caldes, G.L. Mutter, P. Platzer, and C. Eng. "Breast-cancer stromal cells with TP53 mutations and nodal metastases". In: *New England Journal of Medicine* 357.25 (2007), pp. 2543–2551. URL: http://www.nejm.org/doi/full/10.1056/NEJMoa071825 (see p. 31).

[62] Dirk Repsilber, Sabine Kern, Anna Telaar, Gerhard Walzl, Gillian F Black, Joachim Selbig, Shreemanta K Parida, Stefan H E Kaufmann, and Marc Jacobsen. "Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach." In: *BMC bioinformatics* 11 (2010), p. 27. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-27. URL: http://www.ncbi.nlm.nih.gov/pubmed/20070912 (see pp. 3, 5, 6, 8, 10, 22, 25, 26, 30, 32).

[63] Sushmita Roy, Terran Lane, Chris Allen, Anthony D Aragon, and Margaret Werner-Washburne. "A hidden-state Markov model for cell population deconvolution." In: *Journal of computational biology : a journal of computational molecular cell biology* 13.10 (2006), pp. 1749–74. ISSN: 1066-5277. DOI: 10.1089/cmb.2006.13.1749. URL: http://www.ncbi.nlm.nih.gov/pubmed/17238843 (see p. 3).

[64] Stephanie Schneider, Temple Smith, and Ulla Hansen. "SCOREM: statistical consolidation of redundant expression measures." In: *Nucleic acids research* (2011), pp. 1–12. ISSN: 1362-4962. DOI: 10.1093/nar/gkr1270. URL: http://www.ncbi.nlm.nih.gov/pubmed/22210887 (see p. 31).

[65] U Schneider, H U Schwenk, and G Bornkamm. "Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma." In: *International journal of cancer. Journal international du cancer* 19.5 (1977), pp. 621–6. ISSN: 0020-7136. URL: http://www.ncbi.nlm.nih.gov/pubmed/68013 (see p. 22).

[66] Shai S Shen-Orr, Robert Tibshirani, Purvesh Khatri, Dale L Bodian, Frank Staedtler, Nicholas M Perry, Trevor Hastie, Minnie M Sarwal, Mark M Davis, and Atul J Butte. "Cell type-specific gene expression differences in complex tissues." In: *Nature methods* 7.4 (Apr. 2010), pp. 287–9. ISSN: 1548-7105. DOI: 10.1038/nmeth.1439. URL: http://www.ncbi.nlm.nih.gov/pubmed/20208531 (see pp. 3–5, 7, 10, 12, 13, 18, 22, 23, 26).

[67] Leming Shi et al. "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." In: *Nature biotechnology* 24.9 (Sept. 2006), pp. 1151–61. ISSN: 1087-0156. DOI: 10.1038/nbt1239. URL: http://www.ncbi.nlm.nih.gov/pubmed/16964229 (see p. 18).

[68] Sandra Siegert, Erik Cabuy, Brigitte Gross Scherf, Hubertus Kohler, Satchidananda Panda, Yun-Zheng Le, Hans Jörg Fehling, Dimos Gaidatzis, Michael B Stadler, and Botond Roska. "Transcriptional code and disease map for adult retinal cell types." In: *Nature neuroscience* January (Jan. 2012). ISSN: 1546-1726. DOI: 10.1038/nn.3032. URL: http://www.ncbi.nlm.nih.gov/pubmed/22267162 (see p. 18).

[69] Karline Soetaert, Karel Van den Meersche, and Dick van Oevelen. *limSolve: Solving Linear Inverse Models*. R package 1.5.1. 2009 (see p. 29).

[70] Robert O Stuart, William Wachsman, Charles C Berry, Jessica Wang-Rodriguez, Linda Wasserman, Igor Klacansky, Dan Masys, Karen Arden, Steven Goodison, Michael McClelland, Yipeng Wang, Anne Sawyers, Iveta Kalcheva, David Tarin, and Dan Mercola. "In silico dissection of cell-type-associated patterns of gene expression in prostate cancer." In: *Proceedings of the National Academy of Sciences of the United States of America* 101.2 (2004), pp. 615–20. ISSN: 0027-8424. DOI: 10.1073/pnas.2536479100. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=327196&tool=pmcentrez&rendertype=abstract (see p. 3).

[71] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith a Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P Cooke, John R Walker, and John B Hogenesch. "A gene atlas of the mouse and human protein-encoding transcriptomes." In: *Proceedings of the National Academy of Sciences of the United States of America* 101.16 (2004), pp. 6062–7. ISSN: 0027-8424. DOI: 10.1073/pnas.0400782101. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=395923\&tool=pmcentrez\&rendertype=abstract (see p. 11).

[72] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael a Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.43 (2005), pp. 15545–50. ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102. URL: http://www.ncbi.nlm.nih.gov/pubmed/16199517 (see pp. 6, 11, 12).

[73] S Tsuchiya, M Yamabe, Y Yamaguchi, Y Kobayashi, T Konno, and K Tada. "Establishment and characterization of a human acute monocytic leukemia cell line (THP-1)." In: *International journal of cancer. Journal international du cancer* 26.2 (1980), pp. 171–6. ISSN: 0020-7136. URL: http://www.ncbi.nlm.nih.gov/pubmed/6970727 (see p. 22).

[74] K. Van den Meersche, Karline Soetaert, and D. Van Oevelen. "xsample (): An R function for sampling linear inverse problems". In: *Journal of Statistical Software* 30.Code Snippet 1 (2009), pp. 1–15. URL: http://biblio.ugent.be/input/download?func=downloadFile\&fileOId=677001 (see p. 29).

[75] D Venet, F Pecasse, C Maenhaut, and H Bersini. "Separation of samples into their constituents using gene expression data". In: *Bioinformatics* 17.suppl 1 (2001), S279. URL: http://bioinformatics.oxfordjournals.org/content/17/suppl_1/S279.short (see pp. 3, 5, 6).

[76] Min Wang, Stephen R Master, and Lewis a Chodosh. "Computational expression deconvolution in a complex mammalian organ." In: *BMC bioinformatics* 7 (2006), p. 328. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-328. URL: http://www.ncbi.nlm.nih.gov/pubmed/16817968 (see pp. 3, 5, 8, 10).

[77] Nicholas a Watkins et al. "A HaemAtlas: characterizing gene expression in differentiated human blood cells." In: *Blood* 113.19 (2009), e1–9. ISSN: 1528-0020. DOI: 10.1182/blood-2008-06-162958. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2680378\&tool=pmcentrez\&rendertype=abstract (see pp. 16, 18–24, 29, 41, 43).

[78] A.R. Whitney, M Diehn, S.J. Popper, A.A. Alizadeh, J.C. Boldrick, D.A. Relman, and P.O. Brown. "Individuality and variation in gene expression patterns in human blood". In: *Proceedings of the National Academy of Sciences of the United States of America* 100.4 (2003), p. 1896. URL: http://www.pnas.org/content/100/4/1896.short (see p. 18).

[79] Sheng-Jian Xiao, Chi Zhang, Quan Zou, and Zhi-Liang Ji. "TiSGeD: a database for tissue-specific genes." In: *Bioinformatics (Oxford, England)* 26.9 (2010), pp. 1273–5. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq109. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2859128\&tool=pmcentrez\&rendertype=abstract (see p. 11).

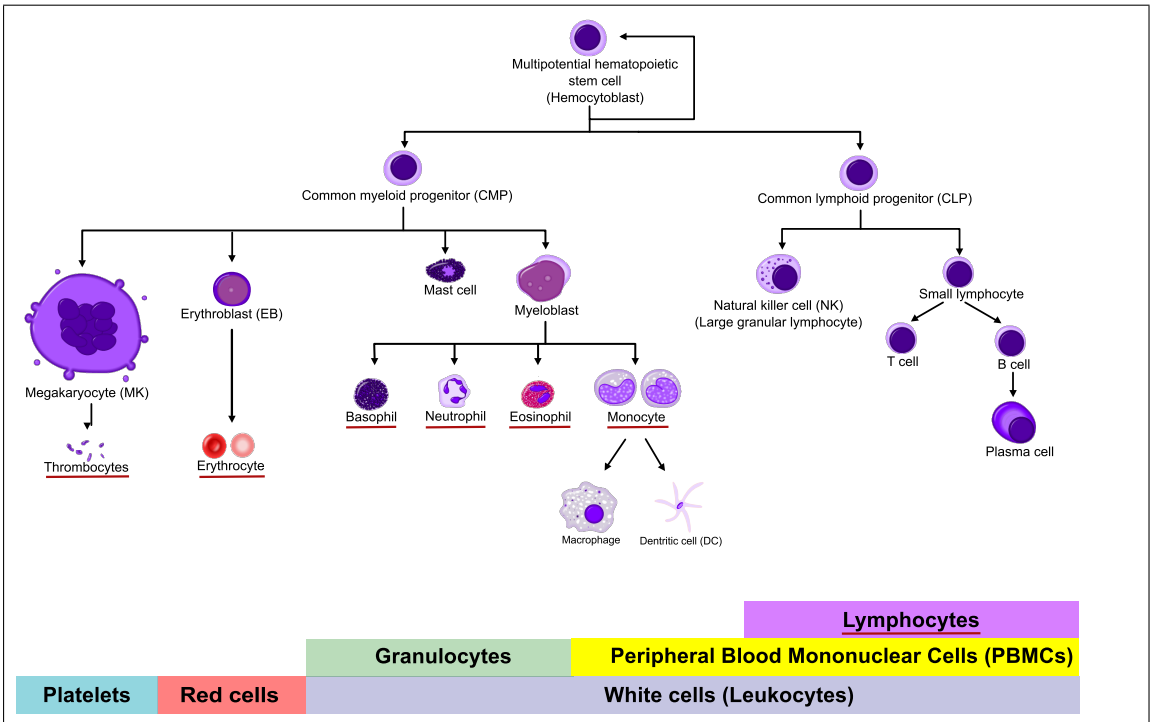
[80] Xiaoqin Yang, Yun Ye, Guiping Wang, Hong Huang, Dekuang Yu, and Shuang Liang. "VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery." In: *Physiological genomics* 43.8 (2011), pp. 457–60. ISSN: 1531-2267. DOI: 10.1152/physiolgenomics.00178.2010. URL: http://physiolgenomics.physiology.org/cgi/content/abstract/43/8/457 (see pp. 11, 18).

[81] Xinan Yang and Xiao Sun. "Meta-analysis of several gene lists for distinct types of cancer: a simple way to reveal common prognostic markers." In: *BMC bioinformatics* 8 (2007), p. 118. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-118. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1853113\&tool=pmcentrez\&rendertype=abstract (see p. 30).

[82] Yingdong Zhao and Richard Simon. "Gene expression deconvolution in clinical samples." In: *Genome medicine* 2.12 (2010), p. 93. ISSN: 1756-994X. DOI: 10.1186/gm214. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3025435\&tool=pmcentrez\&rendertype=abstract (see p. 3).

[83] Yi Zhong, Yin-Wooi Wan, Kaifang Pang, Lionel Ml Chow, and Zhandong Liu. "Digital sorting of complex tissues for cell type-specific gene expression profiles". In: *BMC Bioinformatics* 14.1 (2013), p. 89. ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-89. URL: http://www.biomedcentral.com/1471-2105/14/89 (see p. 26).

A Hematopoietic tree



Modified version from an image from the public domain[38]

Figure 14: *Hematopoietic tree. Schematic representation of the main lineages of blood cells. The multipotential hematopoietic stem cell (Hemmocytoblast), differentiates into the common lymphoid progenitor (CLP) and common myeloid progenitor (CMP), from which arise the lymphoid and myeloid lineages respectively. CLPs are progenitors for all lymphocytes, further differentiating into Natural killer cells (NK), and B and T cells respectively. CMPs give rise to megakaryocytes (MK) and erythroblasts (EB) , that are progenitors for thrombocytes (platelets) and erythrocytes (red cells) respectively, and to myeloblasts that are progenitors for granulocytes and monocytes. Monocytes differentiate in tissues into macrophages and dendritic cells (DC). Some B cells differentiate on activation into plasma cells, i.e. effector B cells, that produce large amount of pathogen-specific antibodies, which scale up immune response and target pathogens (Ferreira et al. 2005; Janeway et al. 2001). Cell types that are underlined in red are those for which proportions are available in CBC data.*

[38] http://commons.wikimedia.org/wiki/File:Hematopoiesis_simple.svg published under the terms of the GNU Free Documentation License.

B Marker composition

```
# For Human (Homo Sapiens)
(([hs_tissue_distribution-TissueType:*] & [hs_tissue_distribution-TissueLevel#5:5])
& [hs_tissue_distribution-TissueRank#1:1])

# For Rat (Rattus Norvegicus)
(([rn_tissue_distribution-TissueType:*] & [rn_tissue_distribution-TissueLevel#5:5])
& [rn_tissue_distribution-TissueRank#1:1])
```

Figure 15: *Queries used to extract the most tissue-specific UniGene EST clusters in Human and Rat from the TissueDistributionDB database. We selected only the EST clusters that ranked first in terms of tissue-specific expression (Rank 1) in the tissues defined by the more detailed tissue classification (Level 5).*

C Marker IDs conversion pipeline

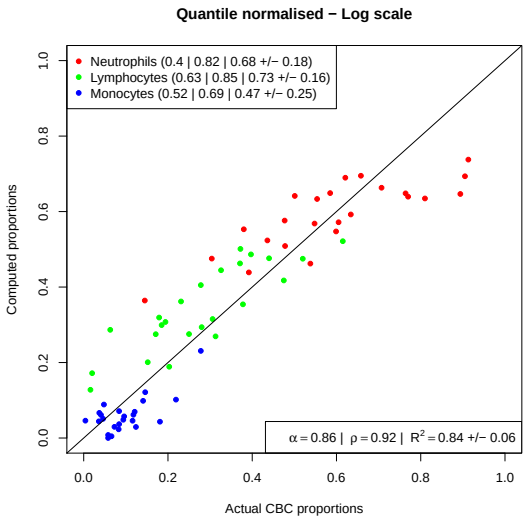
```
# load data from registry
m <- MarkerList("HaemAtlas")
e <- ExpressionMix("GSE11058")

# convert Affy IDs from HGU133A/B to HGUplus2 (showing details of the
# mapping process)
m_affy_all <- convertIDs(m, e, verbose = 4)

## # Converting 2069 markers from Annotation (illuminaHumanv2.db) to Annotation (hgu133plus2.db) ...
## # Converting from Annotation (illuminaHumanv2.db) to Annotation (hgu133plus2.db) ...
## # Limiting query to Annotation (illuminaHumanv2.db) ... [2069 -> 2069 id(s)]
##  # Loading map(s) from Annotation (illuminaHumanv2.db) to Annotation (hgu133plus2.db)  [x-platform /x-probe id] ...  OK [2 step(s)]
##  # Mapping from Annotation (illuminaHumanv2.db) to EntrezId (illuminaHumanv2.db) [49481 entries] ...  [1803/2069 mapped (1:1)]
##  # Applying filtering strategy 'auto' ...  [1803/1803 passed (1:1)]
##  # Mapping from EntrezId (hgu133plus2.db) to Annotation (hgu133plus2.db) [43059 entries] ...  [1737/1803 mapped (1:1-14 = 4170)]
##  # Applying filtering strategy 'auto' ... (kept 1703 2nd-affy probes) (trunking 1088 maps to 1:1)  [1737/1737 passed (1:1)]
## OK [1737/2069 mapped (1:1)]
## OK [1737/2069 (1:1)]
## # Processing 2069 markers from Annotation (illuminaHumanv2.db) to Annotation (hgu133plus2.db) ...
##  ** Processing ids for 'B-CD19' ...
##  # Removing duplicated id(s) ...  OK [dropped 11/187 id(s)]
##  OK [176/247 (1:1)]
##  ** Processing ids for 'Erythroblast' ...
##  # Removing duplicated id(s) ...  OK [dropped 11/291 id(s)]
##  OK [280/322 (1:1)]
##  ** Processing ids for 'Granulocyte-CD66b' ...
##  # Removing duplicated id(s) ...  OK [dropped 33/702 id(s)]
##  OK [669/878 (1:1)]
##  ** Processing ids for 'Megakaryocyte' ...
##  # Removing duplicated id(s) ...  OK [dropped 8/258 id(s)]
##  OK [250/279 (1:1)]
##  ** Processing ids for 'Monocyte-CD14' ...
##  # Removing duplicated id(s) ...  OK [dropped 7/188 id(s)]
##  OK [181/205 (1:1)]
##  ** Processing ids for 'NK-CD56' ...
##  # Removing duplicated id(s) ...  OK [dropped 3/65 id(s)]
##  OK [62/82 (1:1)]
##  ** Processing ids for 'T-CD4' ...
##  # Removing duplicated id(s) ...  OK [dropped 1/42 id(s)]
##  OK [41/51 (1:1)]
##  ** Processing ids for 'T-CD8' ...  OK [4/5 (1:1)]
## # Checking for duplicated marker(s) across cell-types ... OK [dropped 4/1663]
## OK [1659/2069 (1:1)]
```

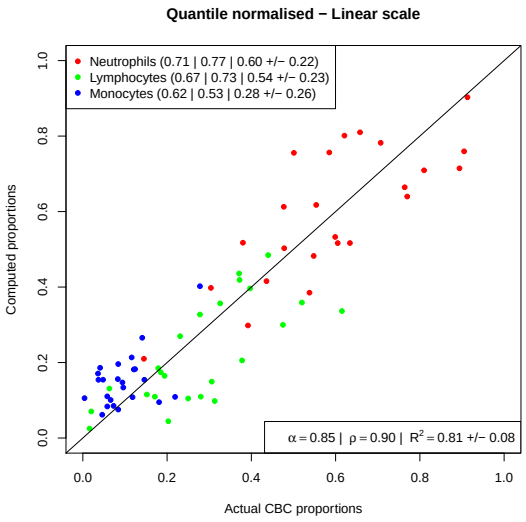
Figure 16: *Sample code for converting the HaemAtlas marker list defined by Watkins et al. (2009), from Illumina HumanWG6v2 probe identifiers to Affymetrix HG U133 Plus 2.0 probeset ids. This code is identical to the one shown in Figure 8, but uses verbose output, that shows more details on how the mapping is performed.*

D Effect of scaling and normalization method

```
# Quantile normalized - Log-scale
res <- gedBlood(e)
```

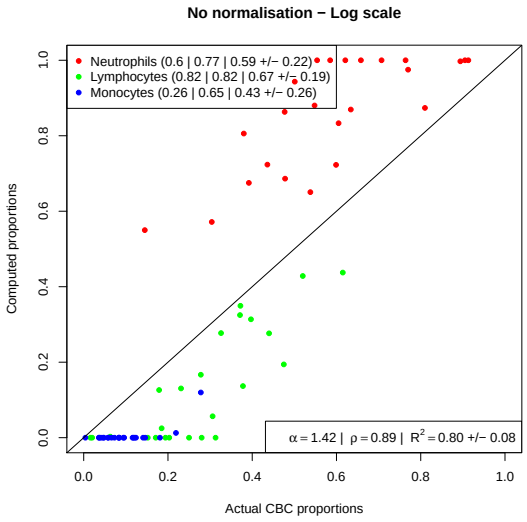
```
# Quantile normalized - Linear scale
res <- gedBlood(expb(e, 2))
```






```
# Log-transformed only
res <- gedBlood(e, normalize = FALSE)
```

```
# Linear scale
res <- gedBlood(expb(e, 2), normalize = FALSE)
```



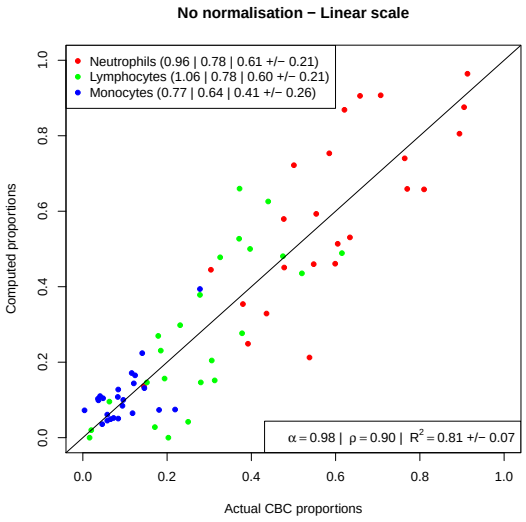


Figure 17: *Computed proportions vs. Actual CBC proportions of lymphocytes, monocytes and neutrophils for dataset* GSE20300. *The deconvolution is performed using the* `qprog` *algorithm in combination with the basis signature matrix from Abbas et al. (2009). Computed proportions are obtained from the aggregation of proportions of the detailed subset of immune cells into their respective category.*

E Assessing marker expressions in pure samples

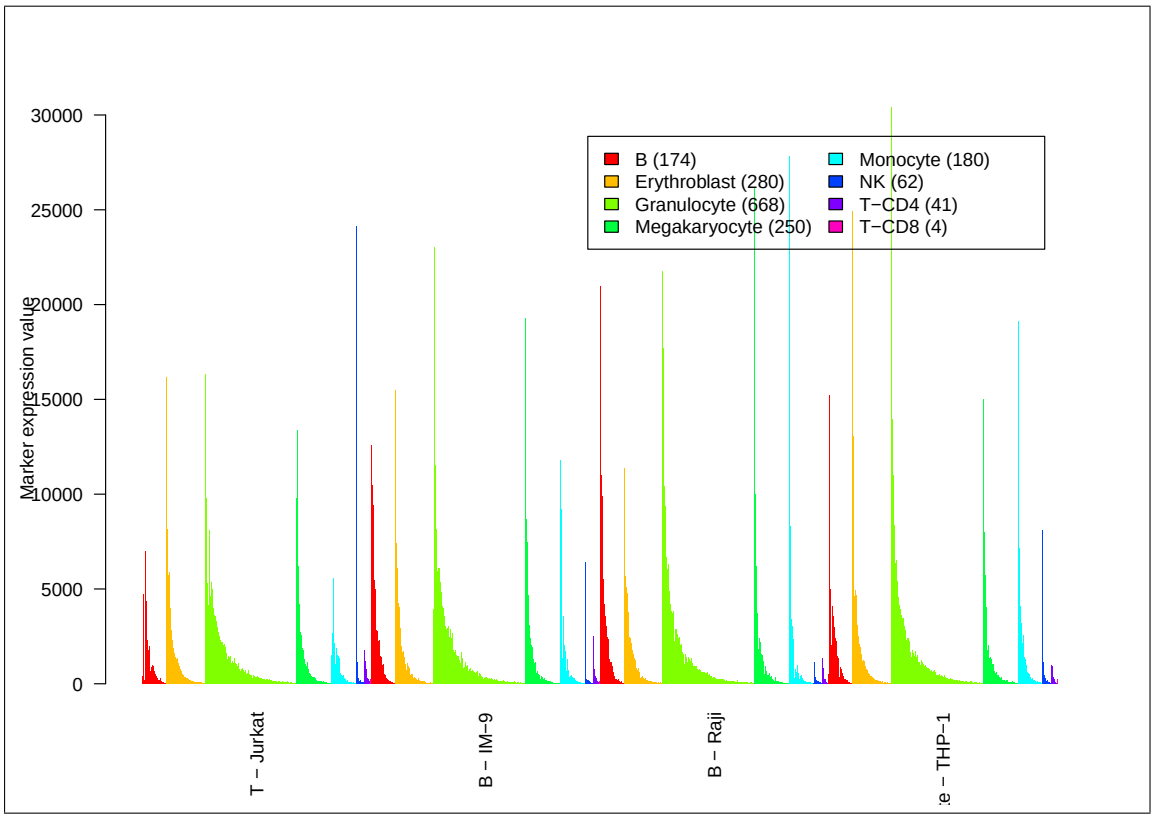


Figure 18: *Expression values of all the marker genes of each cell type in the list from HaemAtlas (Watkins et al. 2009), ordered by decreasing uniform expression in each immune cell line from dataset* `GSE11058`*. This figure is the full version of Figure 10, where only the top 50 markers of each cell type were shown.*