CellMix: A Comprehensive Toolbox for Gene Expression Deconvolution

Renaud Gaujoux¹, Cathal Seoighe²*

¹Computational Biology Group, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, South Africa
²National University of Ireland Galway, Ireland

ABSTRACT

Summary: Gene expression data are typically generated from heterogeneous biological samples that are composed of multiple cell or tissue types, in varying proportions, each contributing to global gene expression. This heterogeneity is a major confounder in standard analysis such as differential expression analysis, where differences in the relative proportions of the constituent cells may prevent or bias the detection of cell-specific differences. Computational deconvolution of global gene expression is an appealing alternative to costly physical sample separation techniques, and enables a more detailed analysis of the underlying biological processes, at the cell type level. To facilitate and popularise the application of such methods, we developed CellMix, an R package that incorporates most state of the art deconvolution methods, into an intuitive and extendible framework, providing a single entry point to explore, assess and disentangle gene expression data from heterogeneous samples.

Availability and Implementation: The CellMix package builds upon R/BioConductor and is available from http://web.cbio.uct.ac.za/~renaud/CRAN/web/CellMix. It is currently being submitted to BioConductor. The package’s vignettes notably contain additional information, examples and references.

Contact: renaud@cbio.uct.ac.za

1 GENE EXPRESSION DECONVOLUTION

The vast majority of gene expression data are generated from biological samples that are composed of multiple cell or tissue types that contribute to different extents to the global gene expression, according to their relative proportions. Heterogeneity in sample composition is commonly acknowledged as a major confounder in classical gene expression analysis like differential expression analysis, specially in clinical studies (Zha and Simon, 2010). In this context, being able to disentangle the effects due to cell-specific expression and/or varying proportions provides finer insights into the biological processes of interest, by enabling the data to be explored at the cell type level.

Gene expression deconvolution receives constant interest in bioinformatics research, with new methodologies published regularly (Zha and Simon, 2010). While all methods apply to global expression data, they differ in the type of auxiliary data they required, such as cell proportion measurements/estimates, cell-specific signatures or sets of marker genes. Having a standardised and unified interface for running a variety of deconvolution methods that can adapt to most common data settings, would therefore be very useful, and help popularise computational deconvolution.

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

This paper briefly describes the main features of the CellMix package, and illustrates its capability with some concrete examples. More examples, as well as thorough documentation, references and implementation details are available in the package’s vignettes.

2 THE CELLMIX PACKAGE: OVERVIEW

The CellMix package builds upon the Bioconductor project (Gentleman et al., 2004) and the NMF package (Gaujoux and Seoighe, 2010), to provide a flexible general framework for gene expression deconvolution methods. It defines a rich programming interface around three internal extendible registries dedicated to deconvolution methods, marker gene lists and benchmark datasets, respectively.

2.1 Deconvolution methods

CellMix provides access to a range of 7 gene expression deconvolution methods, in such a way that they can easily be applied to commonly available data, via a unique interface function called ged. In particular, we implemented a default method selection scheme, which chooses a sensible deconvolution method based on the type of input and auxiliary data that are provided (See section Algorithms on the package’s webpage for details on each available method).

2.2 Cell signatures and marker gene sets

In the context of gene expression deconvolution and sample heterogeneity in general, marker genes constitute a critical asset. For example, they can provide cell-specific signals that can be used to estimate cell-specific signatures and/or cell proportions, or detect cell type-related differential expression (Gaujoux and Seoighe, 2011; Kuhn et al., 2011; Bolen et al., 2011). The CellMix package includes a set of 8 marker gene lists, compiled from previous studies and public databases, and provides many convenient filtering or plotting functions for such type of data. Moreover, it implements

*to whom correspondence should be addressed
a very flexible general pipeline to convert gene identifiers, including across platform or species, which greatly simplifies the use of both marker genes and datasets from one study in another.

2.3 Benchmark datasets

CellMix ships with a curated repository of 11 public datasets compiled from a variety of published studies on cell/tissue specific expression gene or deconvolution methods. These datasets were chosen because they contain not only global gene expression from mixed samples, but also data such as cell type specific signatures and/or measured mixture proportions for each sample, making them ideal for developing and validating deconvolution approaches. Each dataset can be loaded in a single call, which applies a pre-processing pipeline to the original – normalised – data, downloaded from public repositories. In particular, data relevant for deconvolution are extracted from sample annotations and processed into a single data object, from which mixed/pure sample expression profiles and/or cell proportions can be easily retrieved.

3 EXAMPLE: BLOOD SAMPLE DECONVOLUTION

The dataset GSE20300 contains gene expression data (on Affymetrix HGU133Plus2) of whole blood samples from stable and acute rejection pediatric kidney transplant, for which Complete Blood Count (CBC) data are available (Shen-Orr et al., 2010). The following code estimates these proportions using an optimised set of immune cell type signatures (on Affymetrix HGU133A/B) (Abbas et al., 2009) and produces the scatter plot in Figure 1. Both these data are available in the CellMix package. Importantly, sensible probeset mapping or joint data transformation and normalisation are transparently handled via a – customisable – pre-processing pipeline. To our knowledge, this is the first time these data have been used in this way. The ease with which the results are generated highlights the usefulness of the CellMix package.

```r
# load benchmark data
acr <- ExpressionMix("GSE20300")
# compute proportions
res <- gedBlood(acr)
# plot against actual CBC
profplot(acr, asCBC(res))
```

4 EXAMPLE: WORKING WITH MARKER GENES

In this example, we illustrate how CellMix simplifies working with marker genes lists. The consistency of expression profiles from 4 transformed immune cell lines contained in dataset GSE11058 (on Affymetrix HGU133Plus2) (Abbas et al., 2009) is graphically assessed using the marker gene list from HaemAtlas, which contains markers for 8 immune cell types derived by Watkins et al. (2009) in an independent study (on Illumina Human V2). The following code generates the heatmap in Figure 1. This shows the expression profiles of the HaemAtlas marker genes that are most highly expressed in the cell lines on the columns (50 marker genes per cell type). Rows are scaled into relative expression separately (i.e. sum up to one). The row annotation columns on the right hand side highlight the cell line in which each marker is expressed at the highest level. They show that some markers are highly expressed by cell types other than their own, which suggests either an altered expression profile of these cell lines, or an inadequacy of these markers for this particular dataset.

```r
# load/extract expression data
pure <- pureSamples(ExpressionMix("GSE11058"))
# load/convert HaemAtlas markers
```

![Figure 1](http://bioinformatics.oxfordjournals.org/)

**REFERENCES**


