DeconRNASeq: A Statistical Framework for Deconvolution of Heterogeneous Tissue Samples Based on mRNA-Seq data

Ting Gong, Joseph D. Szustakowski

April 24, 2017

1 Introduction

Heterogeneous tissues are frequently collected (e.g. blood, tumor etc.) from humans or model animals. Therefore mRNA-Seq samples are often heterogeneous with regard to those cell types, which render it difficult to distinguish whether gene expression variation reflects a shift in cell populations, a change of cell-type-specific expression, or both ([1]).

In this vignette, we present an efficient pipeline and methodology: DeconRNASeq, an R package for deconvolution of heterogeneous tissues based on mRNA-Seq data. It adopts a globally optimized non-negative decomposition algorithm through quadratic programming for estimating the mixing proportions of distinctive tissue types in next generation sequencing data. We demonstrate the feasibility and validity of DeconRNASeq across a range of mixing levels and sources using mRNA-Seq data mixed *in-silico* at known concentrations.

We presented the workflow of DeconRNASeq package in this vignette. This tool allows processing of sequencing data for assessing the performance of linear models and estimating accurately mixing fractions for multiple species of tissues or cells, and is even able to provide accurate estimates for relatively rare cell types (<= 0.02). We applied our approach to a realistic simulations involving complex mixtures of multiple tissues derived from an appropriate experimental design.

2 File Structure Requirements

A single study analysis requires at least two inputs:

datasets An m-by-n matrix of expression values, where m is the number of genes (or probes) under consideration and n is the total number of mRNA-Seq mixing samples consisting of multiple samples. These expression values should be normalized in some manner. But we leave the users to select their preferred normalization methods. It should be noted, expression is generally not on the log_2 scale, which may destroy the linear model in the context of mRNA-Seq expression deconvolution.

signatures An m-by-n matrix of expression values, where m is the number of genes (or probes) which are cell- or tissue-type specific and n is the total number of cell or tussi types. These espression values should be normalized in the same manner of datasets. We do not suggest to take log transformation also.

Our demo uses a simulated example data set, which can be accessed using the code given below.

```
> library(DeconRNASeq)
> ## multi_tissue: expression profiles for 10 mixing samples from
> ## multiple tissues
> data(multi_tissue)
> datasets <- x.data[,2:11]
> signatures <- x.signature.filtered.optimal[,2:6]
> proportions <- fraction
```

For the mixtures, there are 28745 genes. And we have 10 samples. *In silico* mixed data were simulated using ([2]) data, with disparate proportions drawn from random numbers. The mixing proportions used by each type of tissue are shown in the following. It should also be noted that we investigated the influence of extremely low numbers of contaminating cell types (<2 percent).

```
> proportions

               brain muscle   lung  liver  heart
reads.1.RPKM  0.0463 0.0323 0.0805 0.0747 0.7662
reads.2.RPKM  0.0606 0.1156 0.0278 0.6960 0.1000
reads.3.RPKM  0.0728 0.6058 0.1051 0.1262 0.0900
reads.4.RPKM  0.0709 0.0887 0.7242 0.0975 0.0188
reads.5.RPKM  0.6672 0.1347 0.0486 0.0674 0.0821
reads.6.RPKM  0.1368 0.2181 0.1764 0.3678 0.1010
reads.7.RPKM  0.0780 0.2100 0.2800 0.1603 0.2717
reads.8.RPKM  0.1250 0.3997 0.1830 0.1198 0.1726
reads.9.RPKM  0.2309 0.1230 0.5723 0.0214 0.0524
reads.10.RPKM 0.4284 0.3242 0.0913 0.0644 0.0917
```

We adopted mRNA-Seq data from the Illumina BodyMap 2.0 (GSE 30611) as a training data set to define tissue-specific signatures for different human tissues (adrenal gland, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells). We selected the overlapped tissues with mixed data and generated tissue-specific transcriptional profiles

We assessed potential expression signatures via the methods described in ([3]) as the basis matrix for deconvolution. In this case study, we conjecture that genes with extremely small or large read counts somehow violate our assumptions of linearity or are otherwise unreliable. Therefore, we putatively removed the genes with RPKM less than 200 within any of the five tissues from our gene signatures. Consequently, for the tissue-type selected gene signatures, we selected 1570 genes that consist of the signatures for the five tissues and deconvoluted the data.

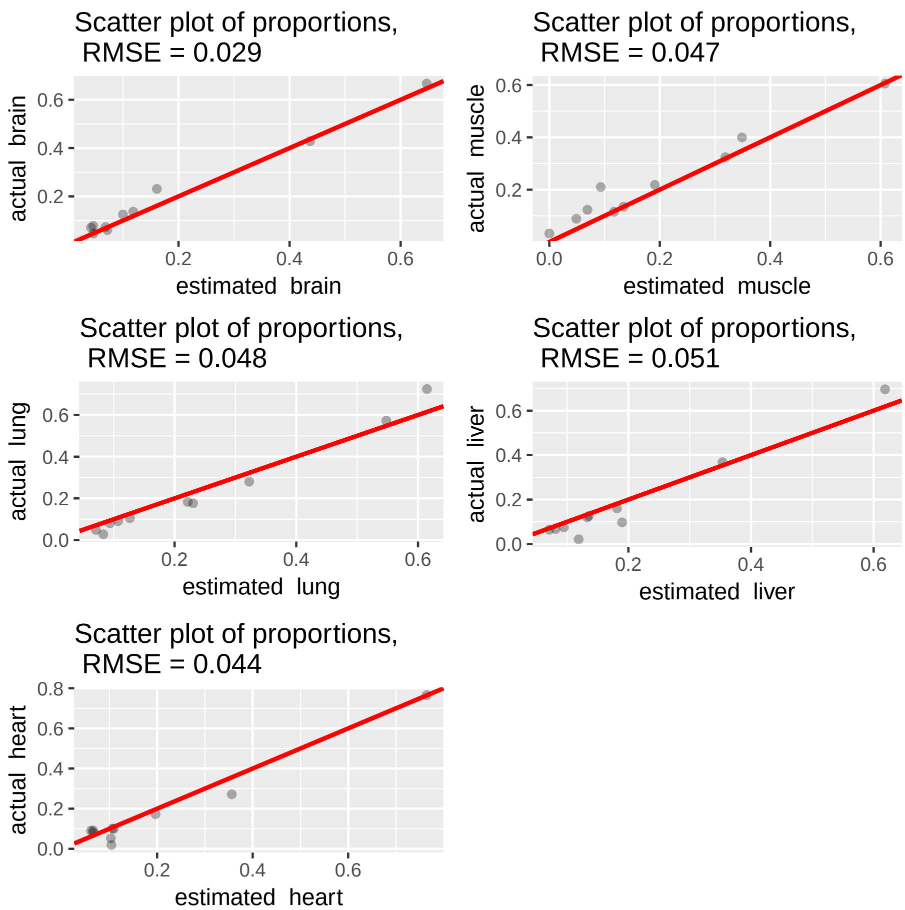
```
> signatures <- x.signature.filtered.optimal[,2:6]
> attributes(signatures)[c(1,2)]

$names
[1] "brain"  "muscle" "lung"   "liver"  "heart"

$class
[1] "data.frame"
```

3 Deconvolution Analysis

After we initiated the parameters/arguments in the last section, we can perform the deconvolution analysis as following.

```
> DeconRNASeq(datasets, signatures, proportions, checksig=FALSE,
+                         known.prop = TRUE, use.scale = TRUE, fig = TRUE)

svd calculated PCA
Importance of component(s):
                 PC1    PC2     PC3    PC4     PC5
R2            0.8388 0.1155 0.02277 0.0177 0.00514
Cumulative R2 0.8388 0.9544 0.97714 0.9948 0.99998

 Attention: the number of pure cell types = 5  defined in the signature matrix;

 PCA results indicate that the number of cell types in the mixtures = 4
$out.all
           brain     muscle       lung      liver
 [1,] 0.04684774 0.00000000 0.09392811 0.09471810
 [2,] 0.07181238 0.11724299 0.08290500 0.61870050
 [3,] 0.06846447 0.60805500 0.12654897 0.13545429
 [4,] 0.04241299 0.04891053 0.61470260 0.18971422
 [5,] 0.64705933 0.13437755 0.07087937 0.08121135
 [6,] 0.11852187 0.19122874 0.23043088 0.35351462
 [7,] 0.04647516 0.09320312 0.32264865 0.18147697
 [8,] 0.09992317 0.34888971 0.22163452 0.13292569
 [9,] 0.16108764 0.06903200 0.54792347 0.11886160
[10,] 0.43684023 0.31901296 0.10717750 0.07104230
           heart
 [1,] 0.76450606
 [2,] 0.10933912
 [3,] 0.06147727
 [4,] 0.10425967
 [5,] 0.06647239
 [6,] 0.10630389
 [7,] 0.35619610
 [8,] 0.19662690
 [9,] 0.10309529
[10,] 0.06592701
```

```
$out.pca
                    PC1     PC2     PC3     PC4     PC5
R2              0.83885 0.11552 0.02277 0.01770 0.00514
Cumulative R2   0.83885 0.95437 0.97714 0.99484 0.99998

$out.rmse
[1] 0.04380197
```



The system we describe here assumes that all relevant cell types are accounted for in the cell-specific expression matrix. In reality, this might not always be the case, as complex tissue samples might include unexpected contaminants, or rare cell populations that have not been previously characterized via expression profiling. Several studies have reported that identifying the correct number of transcriptional source signals in complex samples is very challenging. One common approach is to use principle component analysis (PCA) to estimate the number of sources based on their cumulative variance contributions. Our package includes procedures to assist the user in identifying the appropriate number of sources guided by PCA. When the number of pure cell or tissue types defined in the expression signature matrix is inconsistent with the PCA estimation, our package will give the notification. However, we leave the users to decide the constituent signal components in the mixtures.

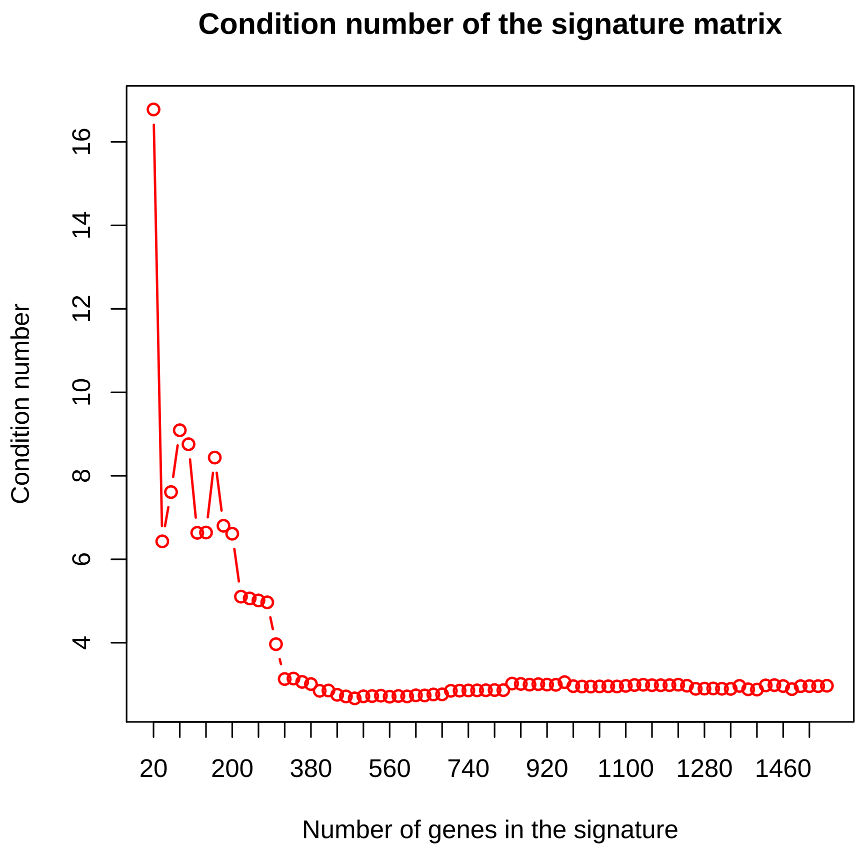
The output out.all includes the estimated mixing fractions of multiple sources in each mixing sample, the out.rmse outputs the mean RMSE(root mean square error) for all estimated tissue proportions if the true proportions are known. We also generated the scatter plots of estimated tissue proportions (y axis) *vs.* actual tissue proportions (x axis) for deconvolution of heterogeneous tissues if fig is TRUE.

4 Condition Number of the Expression Signatures

A basis matrix that contains genes that together form a complete but parsimonious set of robust markers for the tissue types of interest in mRNA-Seq data is crucial to the success of the deconvolution. Therefore, if we know the mixing proportions, a complete set of matrices comprised of different quantities of the most differentially-expressed genes can be tested by comparing the results of each matrix to the known mixture fractions. A matrix' condition number estimates the sensitivity of a system of linear equations to errors in the data. Hence, the condition number is low when the matrix is stable. Thus, we also provide the plot of the condition number *vs.* the number of genes from the gene signature in the deconvolution experiments.

An example is provided by re-runing the same experiment with the parameter checksig to be true.

```
> DeconRNASeq(datasets, signatures, proportions, checksig=TRUE,
+                          known.prop = TRUE, use.scale = TRUE, fig = TRUE)
```



References

[1] Kuhn, A., et al. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain *Nat Meth* 8, 945-947

[2] Pan, Q., et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing *Nat Genet* 40, 1413-1415

[3] Gong, T. Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples *PLoS One* 6, e27156