*Preprint*
xxxxxx
Advance Access Publication Date: 10 May 2019
Application Note

Gene expression

# ADAPTS: Automated Deconvolution Augmentation of Profiles for Tissue Specific cells

**Samuel A Danziger [1]*, David L Gibbs[2], Ilya Shmulevich[2], Mark McConnell [1], Matthew WB Trotter [1,3], Frank Schmitz [1], David J Reiss [1], and Alexander V Ratushny [1]***

[1]Celgene, Seattle WA, 98109, USA; [2]Institute for System Biology, Seattle WA, 98109, USA; [3]Celgene Institute for Translational Research Europe, Sevilla 41092, Spain

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Summary:** Immune cell infiltration of tumors can be an important component for determining patient outcomes, e.g. by inferring immune cell presence by deconvolving gene expression data drawn from a heterogenous mix of cell types. ADAPTS aids deconvolution by adding custom cell types to existing cell-type signature matrices or building new matrices *de novo*. This R package builds a custom signature matrix from purified cell type gene expression data by automatically determining genes that uniquely identify each cell type. The package includes functions that call deconvolution algorithms which use the custom signature matrix to estimate the proportion of cell types present in heterogenous samples.
**Availability:** The R packages ADAPTS, ADAPTSdata, ADAPTSdata2 are at *GitHub* (https://github.com/sdanzige/ADAPTS , etc.) and eventually will be available on *CRAN*
**Contact:** sdanziger@celgene.com, aratushny@celgene.com
**Preprint:** A preprint is available on *BiorXiv* (www.biorxiv.org/content/10.1101/633958v1)

## 1 Introduction

Determining cell type enrichment from gene expression data is an step towards determining tumor immune context (Thorsson *et al*., 2018; Newman *et al*., 2019). One family of techniques for doing this involves regression using a signature matrix (typically with several hundred genes), where each column represents a cell type and each row contains the average gene expression in that cell type (Erkkilä *et al*., 2010; Lähdesmäki *et al*., 2005). These signature matrices are constructed using gene expression from samples of a purified cell type. Generally, the publicly available versions of these gene expression signature matrices use immune cells purified from peripheral blood. Genes are included in these matrices based on how distinct they are for each cell type and how robust the resulting matrix is as measured by matrix stability or prediction accuracy on a test set. Although examples exist of both general purpose immune signature matrices, e.g. LM22 (Newman *et al*., 2015) and Immunostates (Vallania *et al*., 2018), and more tissue specific ones e.g. M17 (Ciavarella *et al*., 2018), these matrices are most likely not appropriate for all diseases and tissue types. One such example would be multiple myeloma whole bone marrow samples, in which both tumor and immune cells are present, immune cells may have different states than in peripheral blood, and non-immune stromal cells such as osteoblasts and adipocytes are expected play an important role in patient outcomes (Bianchi and Munshi, 2015).

One straightforward solution to this problem would be to augment a signature matrix by adding cell types without adding any additional genes. For example, one might find purified adipocyte samples in a public gene expression repository and add the average expression for each gene in the matrix to create an adipocyte augmented signature matrix. While this might work, one might reasonably expect adipocytes to best be identified by genes that are different from those that best characterize leukocytes. We developed a method and an R package ADAPTS (Automated Deconvolution Augmentation of Profiles for Tissue Specific cells) to implement this approach.

## 2 Methods

ADAPTS provides functionality for augmenting an existing cell type signature matrix or even constructing a new signature matrix *de novo*. The user is responsible for finding gene expression data from purified

cells types, such as is available in ArrayExpress (Athar *et al*., 2019) or the Gene Expression Omnibus (Barrett *et al*., 2013). From there, ADAPTS helps a user construct new signature matrices with modular [R] functions and default parameters to:

1. Identify and rank significantly different genes for each cell type.
2. Evaluate the stability (condition number) of many potential solutions.
3. Smooth and normalize to meet tolerances for a robust signature matrix.

Similarly, ADAPTS can be used to construct a *de novo* matrix from first principals rather than starting with a seed matrix. One technique is to build an initial seed matrix out of the $n$ (*e.g.* 100) genes that vary the most between cell types and use ADAPTS to augment that seed matrix. The $n$ initial genes can then be removed from the resulting signature matrix and that new signature matrix can be re-augmented by ADAPTS.

The ADAPTS package includes functionality to call several different deconvolution methods using a common interface, thereby allowing a user to test new signature matrices with multiple algorithms.

These algorithms include:

1. **DCQ** (Altboum *et al*., 2014): An elastic net based deconvolution algorithm that consistently best identifies cell proportions.
2. **SVMDECON** (Newman *et al*., 2015): A support vector machine based deconvolution algorithm.
3. **DeconRNASeq** (Gong *et al*., 2013): A non-negative decomposition based deconvolution algorithm.
4. **Proportions in Admixture** (Langfelder *et al*., 2008): A linear regression based deconvolution algorithm.

## 3 Example: Detecting Tumor Cells

To demonstrate utility of the ADAPTS package, we show how it can be used to augment the LM22 from (Newman *et al*., 2015) to identify myelomatous plasma cells from gene expression profiles of 423 purified tumor ($CD138^+$) samples and 440 whole bone marrow (WBM) samples taken from multiple myeloma patients (Danziger *et al*., 2019). The fraction of myeloma cells, which are tumorous plasma cells, were identified in both sample types via quantification of the cell surface marker CD138. Root mean squared error (RMSE) and Pearson's correlation coefficient ($\rho$) were used to evaluate accuracy of tumor cell fraction estimates. RMSE proved particularly relevant when deconvolving purified $CD138^+$ sample profiles, because 356 of 423 samples are more than 90% pure tumor resulting in clumping of samples with purity near 100%.

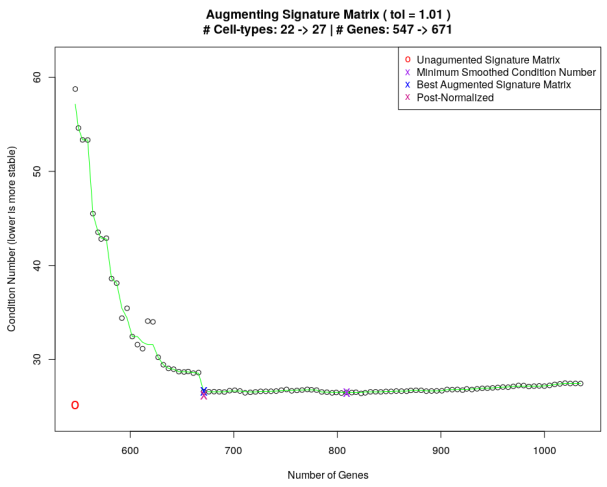


**Fig. 1.** Curve showing the selection of an optimal condition number for MGSM27.

The following matrices were used or generated during the evaluation:

- **LM22**: As reported in (Newman *et al*., 2015). The sum of the 'memory B cells' and 'plasma cells' deconvolved estimates represent tumor percentage.
- **LM22 + 5**: Builds on LM22 by adding purified sample profiles for myeloma specific cell types: plasma memory cells (Mahevas *et al*., 2013), osteoblasts (Athar *et al*., 2019), osteoclasts, adipocytes, and myeloma plasma cells (Torrente *et al*., 2016). The sum estimates for 'memory B cells', 'myeloma plasma cells', 'plasma cells', and 'plasma memory cells' represent tumor percentage.
- **MGSM27**: Builds on LM22 by adding 5 myeloma specific cell types using ADAPTS to determine inclusion of additional genes. Figure 1 shows ADAPTS evaluating matrix stability after adding different numbers of genes, smoothing the condition numbers, and selecting an optimal number of features.
- ***de novo* MGSM27**: Builds a *de novo* MGSM27 from publicly available data similar to those mentioned in (Newman *et al*., 2015) and the 5 myeloma specific cell types.

Table 1 displays average RMSE and $\rho$ for tumor fraction estimates obtained via application of DCQ deconvolution using the four aforementioned matrices across both myeloma profiling datasets.

Table 1. Tumor Identification By DCQ Deconvolution, $N = 10$

| Matrix | WBM *RMSE* | WBM $\rho$ | $CD138^+$ *RMSE* | $CD138^+$ $\rho$ |
|---|---|---|---|---|
| LM22 | $38.0 \pm 0.0$ | $.59 \pm .00$ | $27.0 \pm 0.0$ | $.26 \pm .00$ |
| LM22 + 5 | $37.0 \pm 0.0$ | $.53 \pm .00$ | $12.0 \pm 0.0$ | $.26 \pm .00$ |
| MGSM27 | $23.0 \pm 0.0$ | $.60 \pm .00$ | $09.0 \pm 0.2$ | $.33 \pm .01$ |
| *de novo* MGSM27 | $24.0 \pm 0.0$ | $.65 \pm .00$ | $36.0 \pm 0.0$ | $.18 \pm .00$ |

*RMSE* Root mean square error, $\rho$ Pearson's correlation coefficient, $\pm 1SD$

While the exact genes chosen during each run varies slightly, Table 1 shows that consistently the best accuracy is achieved by augmenting LM22 using ADAPTS. The reduced performance of the *de novo* MGSM27 is likely due to genes that were present in LM22, but were missing in some of the source data and excluded from *de novo* construction. More details are available in the vignette distributed with the R package.

## 4 Conclusion

Table 1 shows an example where including additional genes and tissue specific cell types improves the ability of a deconvolution algorithm to identify tumor fractions in purified and mixed multiple myeloma samples. While this does not demonstrate that the techniques implemented in ADAPTS would be beneficial for all situations, the functions in ADAPTS enable researchers to build their own custom basis matrices to investigate biosamples consisting of multiple cell types.

## Acknowledgements

Thanks to Gareth Morgan, Jake Gockley, Robert Hershberg, Mary H Young, Andrew Dervan, and all other contributors to the paper "Identifying a High-risk Cellular Signature in the Multiple Myeloma Bone Marrow Microenvironment".

## Funding

This work has been supported by the Celgene corporation.

## References

Thorsson, Vésteinn and Gibbs, David L. and Brown, Scott D. and Wolf, Denise and Bortone, Dante S. and Yang, Tai-Hsien Ou and Porta-Pardo, Eduard and Gao, Galen F. and Plaisier, Christopher L. and Eddy, James A. and Ziv, Elad and Culhane, Aedin C. and Paull, Evan O. and Sivakumar, I. K. Ashok and Gentles, Andrew J. and Malhotra, Raunaq and Farshidfar, Farshad and Colaprico, Antonio and Parker, Joel S. and Mose, Lisle E. and Vo, Nam Sy and Liu, Jianfang and Liu, Yuexin and Rader, Janet and Dhankani, Varsha and Reynolds, Sheila M. and Bowlby, Reanne and Califano, Andrea and Cherniack, Andrew D. and Anastassiou, Dimitris and Bedognetti, Davide and Rao, Arvind and Chen, Ken and Krasnitz, Alexander and Hu, Hai and Malta, Tathiane M. and Noushmehr, Houtan and Pedamallu, Chandra Sekhar and Bullman, Susan and Ojesina, Akinyemi I. and Lamb, Andrew and Zhou, Wanding and Shen, Hui and Choueiri, Toni K. and Weinstein, John N. and Guinney, Justin and Saltz, Joel and Holt, Robert A. and Rabkin, Charles E. and Cancer Genome Atlas Research Network and Lazar, Alexander J. and Serody, Jonathan S. and Demicco, Elizabeth G. and Disis, Mary L. and Vincent, Benjamin G. and Shmulevich, Ilya (2018) The Immune Landscape of Cancer, *Immunity*, **4**, 812-830.e14.

Newman, Aaron M. and Steen, Chloé B. and Liu, Chih Long and Gentles, Andrew J. and Chaudhuri, Aadel A. and Scherer, Florian and Khodadoust, Michael S. and Esfahani, Mohammad S. and Luca, Bogdan A. and Steiner, David and Diehn, Maximilian and Alizadeh, Ash A. (2019) Determining cell type abundance and expression from bulk tissues with digital cytometry, *Nature Biotechnology*, 1546-1696.

Erkkilä, Timo and Lehmusvaara, Saara and Ruusuvuori, Pekka and Visakorpi, Tapio and Shmulevich, Ilya and Lähdesmäki, Harr (2010) Probabilistic analysis of gene expression measurements from heterogeneous tissues, *Bioinformatics*, **20**, 2571-2577.

Lähdesmäki, Harri and Shmulevich, Ilya and Dunmire, Valerie and Yli-Harja, Olli and Zhang, Wei (2005) In silico microdissection of microarray data from heterogeneous cell populations, *BMC Bioinformatics*, **1**, 54.

Newman, Aaron M. and Liu, Chih Long and Green, Michael R. and Gentles, Andrew J. and Feng, Weiguo and Xu, Yue and Hoang, Chuong D. and Diehn, Maximilian and Alizadeh, Ash A. (2015) Robust enumeration of cell subsets from tissue expression profiles, *Nature Methods*, **12, 5**, 453-457.

Danziger, Samuel A and McConnell, Mark and Gockley, Jake and Young, Mary H and Ratushny, Alexander V and Morgan, Gareth (2019) Identifying a High-risk Cellular Signature in the Multiple Myeloma Bone Marrow Microenvironment, *bioRxiv*.

Vallania, Francesco and Tam, Andrew and Lofgren, Shane and Schaffert, Steven and Azad, Tej D. and Bongen, Erika and Haynes, Winston and Alsup, Meia and Alonso, Michael and Davis, Mark and Engleman, Edgar and Khatri, Purvesh (2018) Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases, *Nature Communications*, **9**, 4735.

Ciavarella, S. and Vegliante, M. C. and Fabbri, M. and De Summa, S. and Melle, F. and Motta, G. and De Iuliis, V. and Opinto, G. and Enjuanes, A. and Rega, S. and Gulino, A. and Agostinelli, C. and Scattone, A. and Tommasi, S. and Mangia, A. and Mele, F. and Simone, G. and Zito, A. F. and Ingravallo, G. and Vitolo, U. and Chiappella, A. and Tarella, C. and Gianni, A. M. and Rambaldi, A. and Zinzani, P. L. and Casadei, B. and Derenzini, E. and Loseto, G. and Pileri, A. and Tabanelli, V. and Fiori, S. and Rivas-Delgado, A. and López-Guillermo, A. and Venesio, T. and Sapino, A. and Campo, E. and Tripodo, C. and Guarini, A. and Pileri, S. A. (2018) Dissection of DLBCL microenvironment provides a gene expression-based predictor of survival applicable to formalin-fixed paraffin-embedded tissue, *Annals of Oncology*, **12**, 2363-2370.

Bianchi, Giada and Munshi, Nikhil C. (2015) Pathogenesis beyond the cancer clone(s) in multiple myeloma, *Blood*, **12**, 3049-3058.

Altboum, Zeev and Steuerman, Yael and David, Eyal and Barnett-Itzhaki, Zohar and Valadarsky, Liran and Keren-Shaul, Hadas and Meningher, Tal and Mendelson, Ella and Mandelboim, Michal and Gat-Viks, Irit and Amit, Ido (2014) Digital cell quantification identifies global immune cell dynamics during influenza infection, *Molecular Systems Biology*, **10**, 1744-4292.

Gong, Ting and Szustakowski, Joseph D. (2013) DeconRNASeq: A statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data, *Bioinformatics*, **29**, 1367-4803.

Langfelder, Peter and Horvath, Steve (2008) WGCNA: an R package for weighted correlation network analysis, *BMC Bioinformatics*, **9**, 1471-2105.

Athar, A. and Füllgrabe, A. and George, N. and Iqbal, H. and Huerta, L. and Ali, A. and Snow, C. and Fonseca, N. A. and Petryszak, R. and Papatheodorou, I. and Sarkans, U. and Brazma, A. (2019) ArrayExpress update - from bulk to single-cell expression data., *Nucleic acids research*, **47**, D711–D715.

Barrett, Tanya and Wilhite, Stephen E. and Ledoux, Pierre and Evangelista, Carlos and Kim, Irene F. and Tomashevsky, Maxim and Marshall, Kimberly A. and Phillippy, Katherine H. and Sherman, Patti M. and Holko, Michelle and Yefanov, Andrey and Lee, Hyeseung and Zhang, Naigong and Robertson, Cynthia L. and Serova, Nadezhda and Davis, Sean and Soboleva, Alexandra (2013) NCBI GEO: archive for functional genomics data sets—update, *Nucleic acids research*, **41**, D991–D995.

Mahévas, Matthieu and Patin, Pauline and Huetz, François and Descatoire, Marc and Cagnard, Nicolas and Bole-Feysot, Christine and Gallou, Simon Le and Khellaf, Mehdi and Fain, Olivier and Boutboul, David and Galicier, Lionel and Ebbo, Mikael and Lambotte, Olivier and Hamidou, Mohamed and Bierling, Philippe and Godeau, Bertrand and Michel, Marc and Weill, Jean-Claude and Reynaud, Claude-Agnès (2013) B cell depletion in immune thrombocytopenia reveals splenic long-lived plasma cells, *The Journal of Clinical Investigation*, **123**, 432-442.

Torrente, Aurora and Lukk, Margus and Xue, Vincent and Parkinson, Helen and Rung, Johan and Brazma, Alvis (2016) Identification of Cancer Related Genes Using a Comprehensive Map of Human Gene Expression, *PLOS ONE*, **11**, e0157484.