Additional methods

Constrained support Vector Regression

The estimation method proposed by Newman *et al* [1] through Cibersort is based on a mathematical model called Support Vector Regression (SVR). Given a signature matrix and a tumor sample, the algorithm estimates the quantity of the cells present in the tumor sample.
It was already shown that the SVR model is robust to noise and is well suited for the deconvolution task. However, we want to address here two issues of the algorithm proposed in Cibersort.

Cibersort estimates the coefficient using the SVR model and in a first step the algorithm returns coefficients that can be in the whole real numbers set, that is to say even negative values. It means that it is not yet interpreted as proportions coefficients. In their algorithm the authors perform a post normalization process that sets to zero the negative coefficients and then divide by the sum of the coefficients to finally have coefficients that are proportions-like.
We propose to directly address in the model the fact that what is estimated are proportions. This was proposed in the early tries of deconvolution using Linear regression and constrained Linear regression adding first the constraint positivity [2] and then the sum-to-one constraint[3].
Based on this observation, we propose to use the constrained version of the Support Vector Regression which has been proposed in [add ref our paper arxiv: Linear Support Vector Regression with Linear constraints] to estimate the cell proportions.

The constrained SVR has several advantages:

1. The results can directly be interpreted as proportions and do not need post normalization step
2. It is more robust to noise than the classical SVR [add ref our paper arxiv: Linear Support Vector Regression with Linear constraints]
3. It allows more precision in the estimation process as shown in Supplementary Figure 1.

To illustrate these advantages we used the expression microarray dataset from Abbas *et al* [4] (GEO accession number GSE11103) It includes data from four transformed pure cell lines of immune origin: Raji, IM-9 (both from B cells), Jurkat (from T cells), and THP-1 (from monocyte) cells. Three replicates are available for each cell line and corresponding signature matrix was given in Newman *et al.*
We simulated in silico samples by mixing the four different types of cells in several different ratios: each simulated sample was obtained as a linear combination of pure samples expression values weighted by a random proportion. Gaussian noise was added to the data based on the log10 signal to noise ratio (SNR) given by SNR = $10\log_{10}(\frac{\mu^2}{\sigma^2})$, were μ and σ^2 are respectively the mean and the variance of the signal. We chose a SNR value of 15. By this way we generated 100 samples and compared proportions estimated by Cibersort and those estimated using constrained SVR (modified Cibersort). Results are given in Figure 1 above.

As shown on this figure, both methods give equivalent results in terms of correlation to the true proportions when considering all types of cells (higher the best): R=0.91 for Cibersort and R=0.914 for our method (Figures 1A-B). However, we can see that when considering each cell type separately (Figure 1C), constrained SVR (modified Cibersort) leads to better correlations with the true proportions and better RMSE, especially for small proportions. Cibersort algorithm often leads to a zero-coefficient proportion when the cells are present in small quantities whereas constrained SVR gives non-zero results that are closed to the true proportions. The hyperparameters for both models have been chosen using 5-folds cross–validation.

1. Newman A, Liu C, Green M, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*. 2015;12. doi:10.1038/nmeth.3337

2. Qiao W, Quon G, Csaszar E, Yu M, Morris Q, Zandstra PW. PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. *PLOS Computational Biology*. 2012;8(12):1-14. doi:10.1371/journal.pcbi.1002838

3. Gong T, Hartmann N, Kohane IS, et al. Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples. *PLOS ONE*. 2011;6(11):1-11. doi:10.1371/journal.pone.0027156

4. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PLOS ONE*. 2009;4(7):1-16. doi:10.1371/journal.pone.0006098

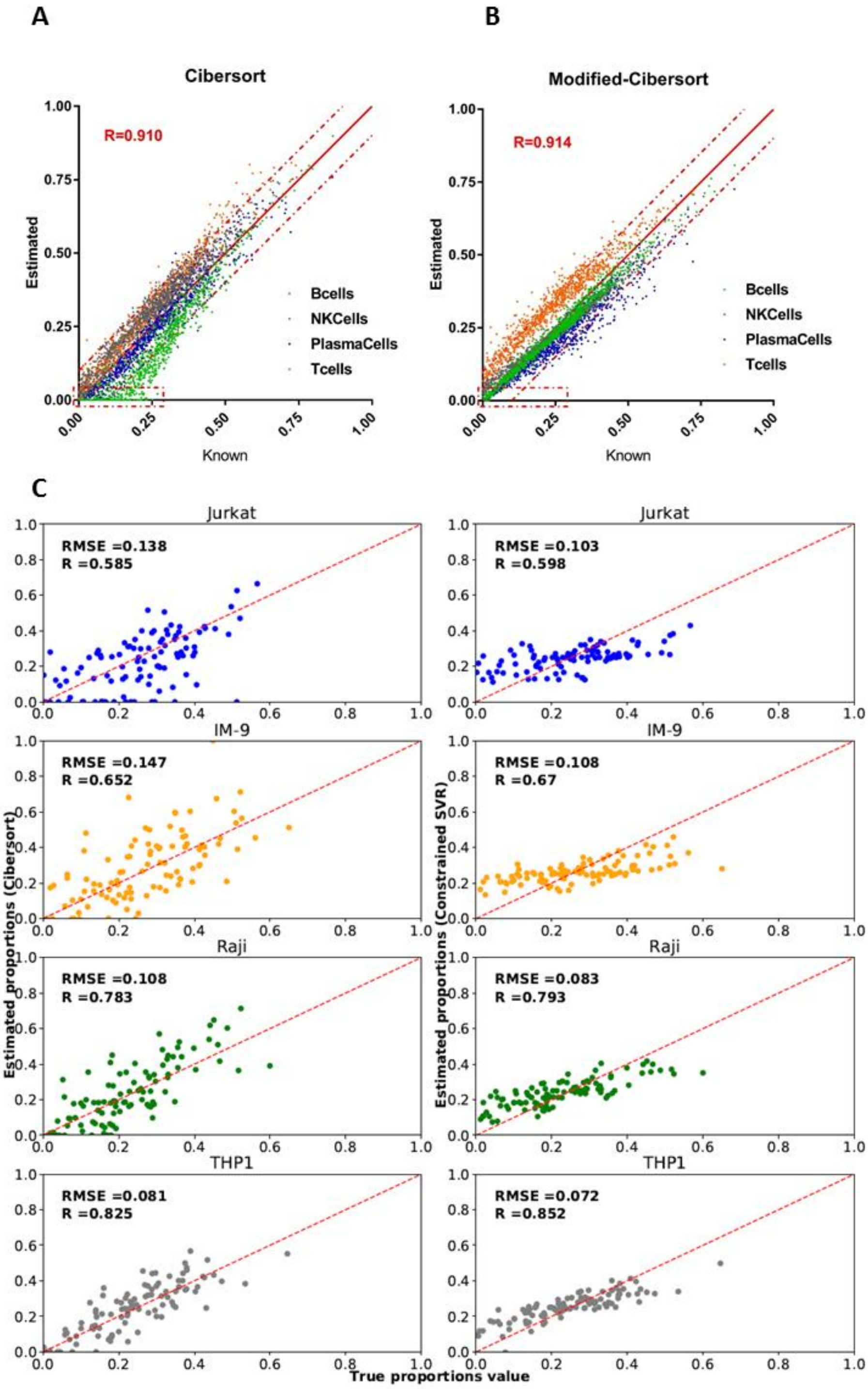

Figure 1: (A-B) Scatter plot representing the estimated proportions of 4 cell subtypes using Cibersort (A) and the modified version of Cibersort (B) as a function of the known proportions. A perfect estimation would represent all the points on the line x=y. The correlation coefficient R is shown as a performance evaluation value. (C) Scatter plots representing proportions of 4 cell subtypes estimated using Cibersort (on the left) or Constrained SVR (on the right) as a function of true proportions values. A perfect estimation would represent all the points on the line x=y depicted in red dashed lines.