Supplementary Material

Evaluation of the pipeline in blood data

In brain, correlation of marker gene profiles can only be indirectly demonstrated due to the unavailability of whole tissue expression data coupled with cell counts. To evaluate the general validity of our entire approach for marker gene selection and MGS summarization as surrogate for cell type proportions, we turned to blood data sets where matching cell count and transcript profiles are available to use as a gold standard. In addition, we compared the MGP-base estimation approach to an alternative method, CIBERSORT (Newman et al., 2015) which is a cell type proportion estimation method that relies on complete expression signatures.

In order to mimic the state of the input data available for human bulk tissue analysis in the brain (namely, mouse-based MGSs) we assembled two matrices of mouse blood cell type expression profiles (mLM22, mLM11), corresponding to the human blood cell type expression profiles used in Newman et al (2015) (hLM22, hLM11). We then applied our marker gene selection pipeline and estimated the MGPs of relevant blood cells in PBMC and LYMPH datasets (see Methods). Interestingly, we found few overlapping genes between mouse and human-based MGS based on our gene selection criteria: mLM22 (707) ∩ hLM22 (228) = 13 genes; mLM11 (705) ∩ hLM11 (161) = 24 genes.

We next examined the correlation of MGP-based estimates with flow cytometry-based cell counts as well as CIBERSORT-based proportion estimations in these datasets. For the LYMPH dataset, flow cytometry data was provided for B-cells and CD4 and CD8 T-cells. Since LM11 was used to predict cellular proportions in the original study, we used mLM11 and hLM11-based MGSs to obtain MGPs. Our estimates were highly correlated to flow cytometry based cell counts and had comparable performance to CIBERSORT (Figure S1A-B).

The PBMC dataset was coupled with counts from eight cell types, represented in the LM22 classification of cell types. We thus used mLM22 and hLM22-based MGSs to calculate MGP for these cell types. Similar to LYMPH dataset analysis, correlation of most MGPs with the actual cell counts was comparable to CIBERSORT-based estimations, with the exception Memory B cells and naïve/resting CD4 T cells and Memory B Cells, where negative correlation was observed between MGP and the actual cell count (Figure S2A-B).

While the exact reason behind the negative correlation is unclear, it might be a result of inter-specie differences or transcriptional changes in lymphoma cells. Different conditions might involve transcriptional changes in some of the marker genes. The impact of such changes on MGP increases with a decrease in the number of marker genes, decreasing the correspondence between MGP and relative cell type proportion. Indeed the two cell types that showed negative correlations had 1 or 2 markers, emphasizing that estimation based on small number of genes should be treated with extreme caution.

Methods

We applied our marker gene discovery method (described in the main paper) to the human blood cell type-specific expression dataset compiled by Newman et al. (2015) which contains 22 cell types and a collapsed list of 11 cell types, referred to by Newman et al. as LM22 and LM11 respectively (here referred to as hLM22 and hLM11 for human). We compiled a similar dataset of mouse cell types by

querying GEO for mouse blood cell types analogous to human cell types in Newman et al. (2015) (**Error! Reference source not found.**). Using the algorithm described above for the neural markers, mouse and human blood cell markers were selected from the relevant set of cell type expression profiles (mLM22 and mLM11 for mouse and hLM22 and hLM11 for human). No marker genes satisfying our criteria were identified for mLM22-based resting CD4 memory T cells and hLM11-based CD4 T cells.

The two bulk tissue datasets used by Newman et al. to estimate cell type profiles (peripheral blood mononuclear cell (PMBC) samples and a lymphoma (LYMPH) cohort) were acquired from cibersort.stanford.edu and GEO (accession: GSE65135) respectively. The corresponding flow cytometry-based cell counts were obtained from cibersort.stanford.edu. The CEL files were preprocessed by the rma function of the affy R package. In order to match the analyses in Newman et al., for the LYMPH dataset, cell type profiles were estimated using marker genes sets of mLM11 and hLM11 while for PMBC dataset estimations were done using the mLM22 and hLM22 marker genes. Profile estimations and CIBERSORT proportion estimations were correlated to flow cytometer based counts using Spearman's ρ.

Figures

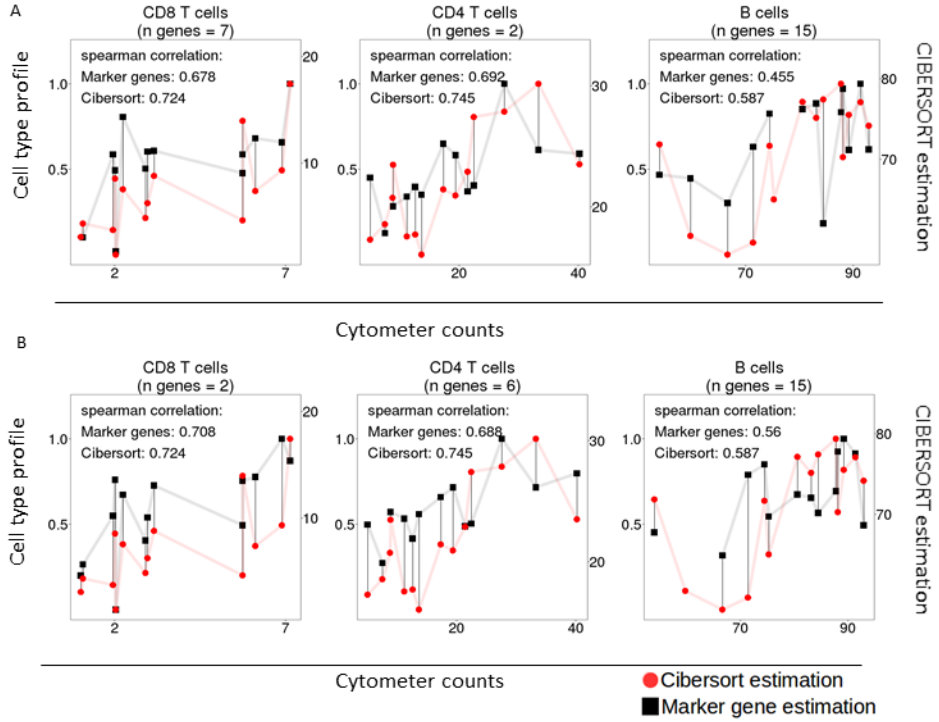

Fig S1: Marker gene profiles are correlated to cell type proportions **(A)** Estimations of cell type profiles (black) and Cibersort estimations(red) are plotted against the cytometer cell counts of the samples. Left axis shows profile estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is a percentage. Estimations done using marker genes selected from human cell type expression profiles based on the hLM11 cell types. (**B)** Estimations done using marker genes selected from mouse cell type expression based on the mLM11 cell types

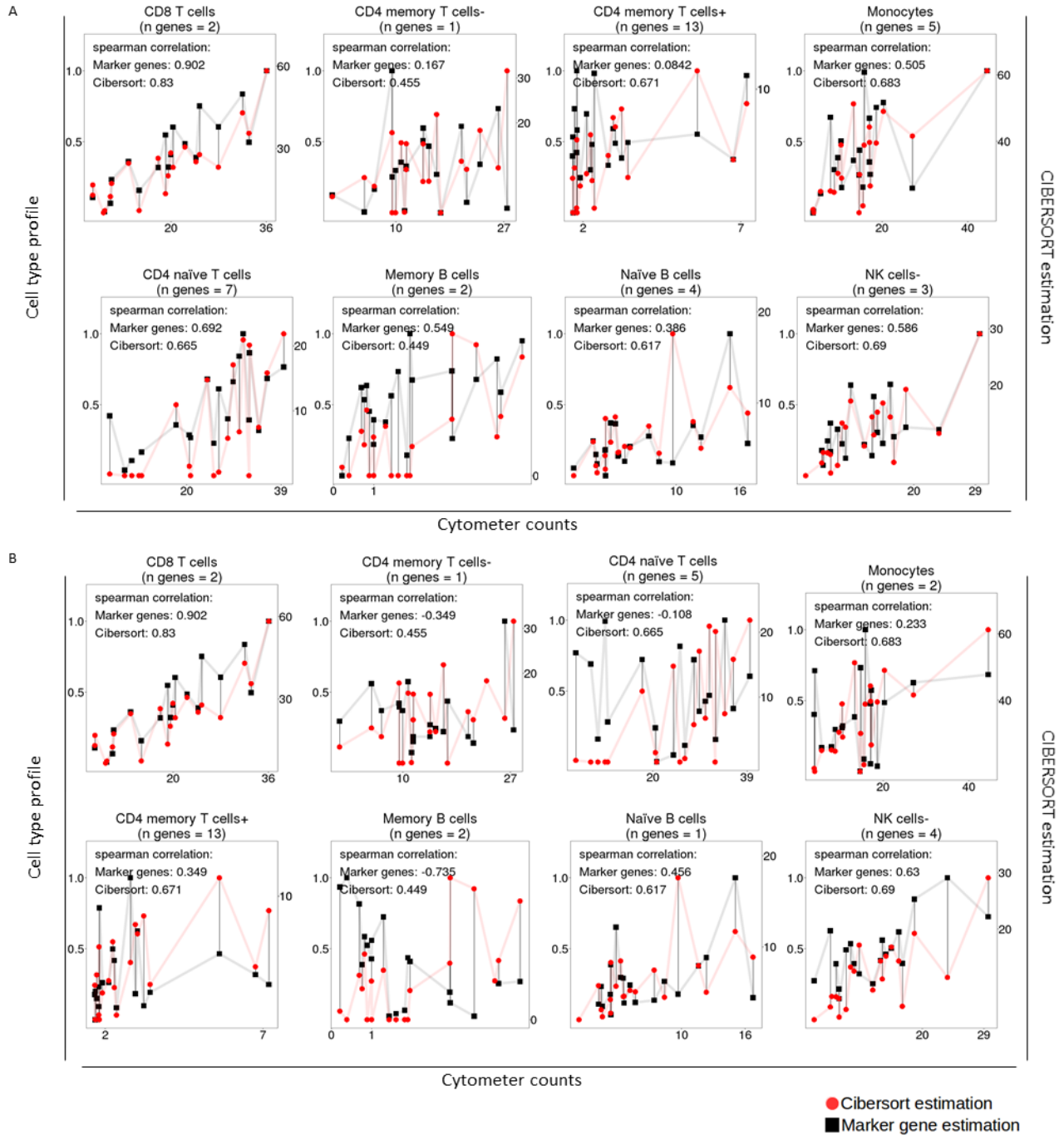

Fig S2: Marker gene profiles are correlated to cell type proportions **(A)** Estimations of cell type profiles (black) and Cibersort estimations(red) are plotted against the cytometer cell counts of the samples. Left axis shows profile estimation values scaled between 0 and 1. Left axis shows Cibersort's estimate which is a percentage. Estimations done using marker genes selected from human cell type expression profiles based on the hLM22 cell types. (**B)** Estimations done using marker genes selected from mouse cell type expression based on the mLM22 cell types