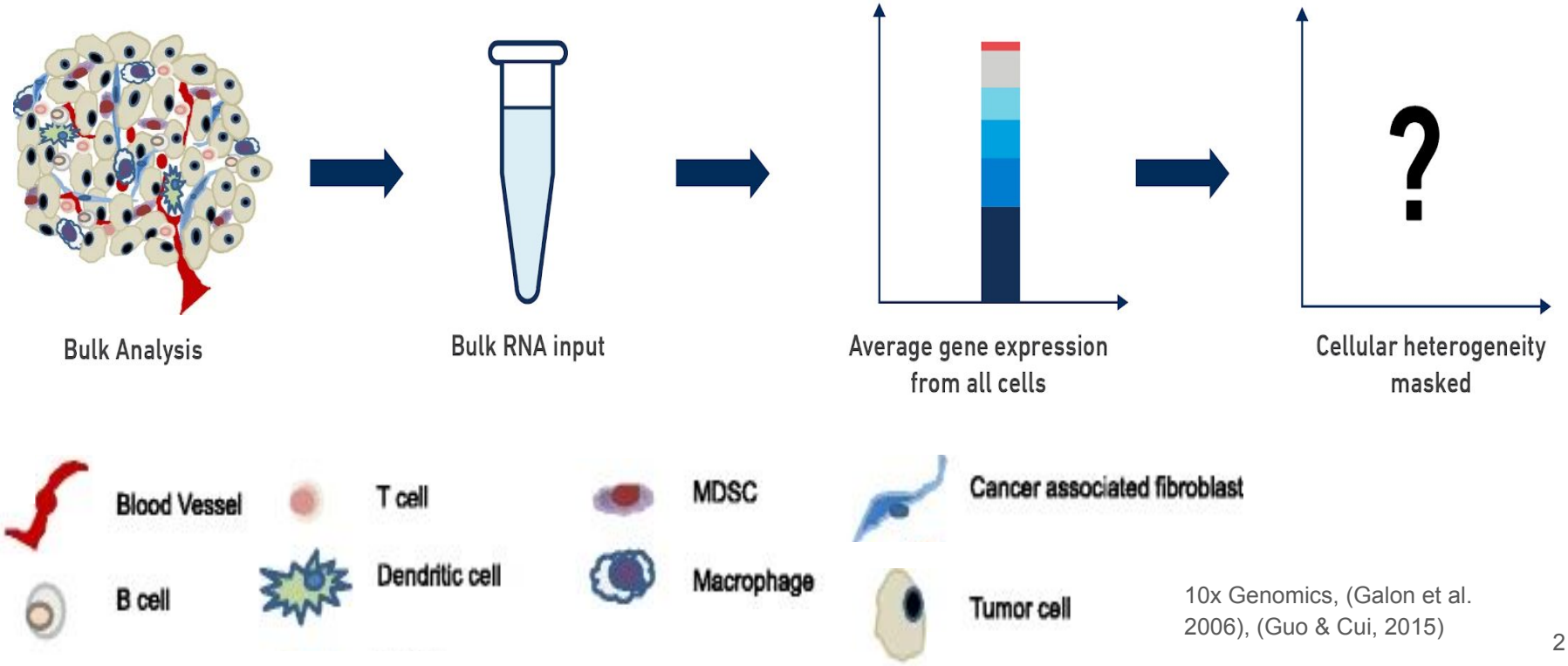
BME 230B: Deconvolution Final Project

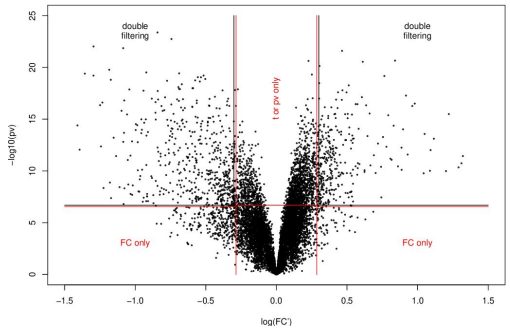
Stephen Hwang

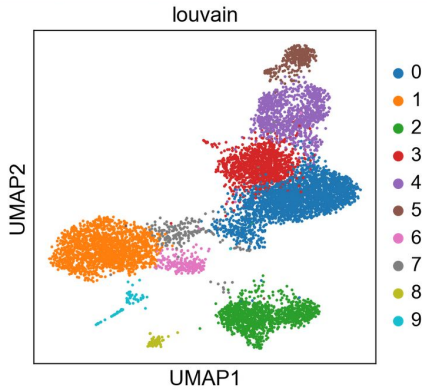
Bulk RNA-seq of the tumor microenvironment masks immune cell type composition predictive of patient outcome

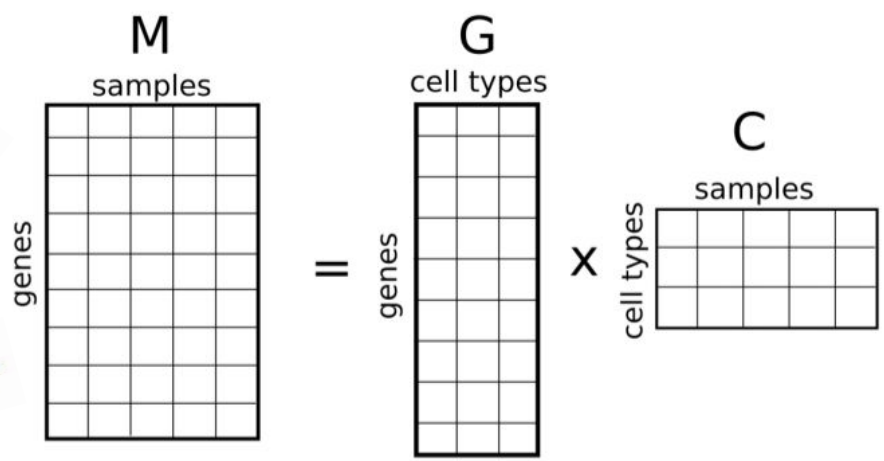


10x Genomics, (Galon et al. 2006), (Guo & Cui, 2015)

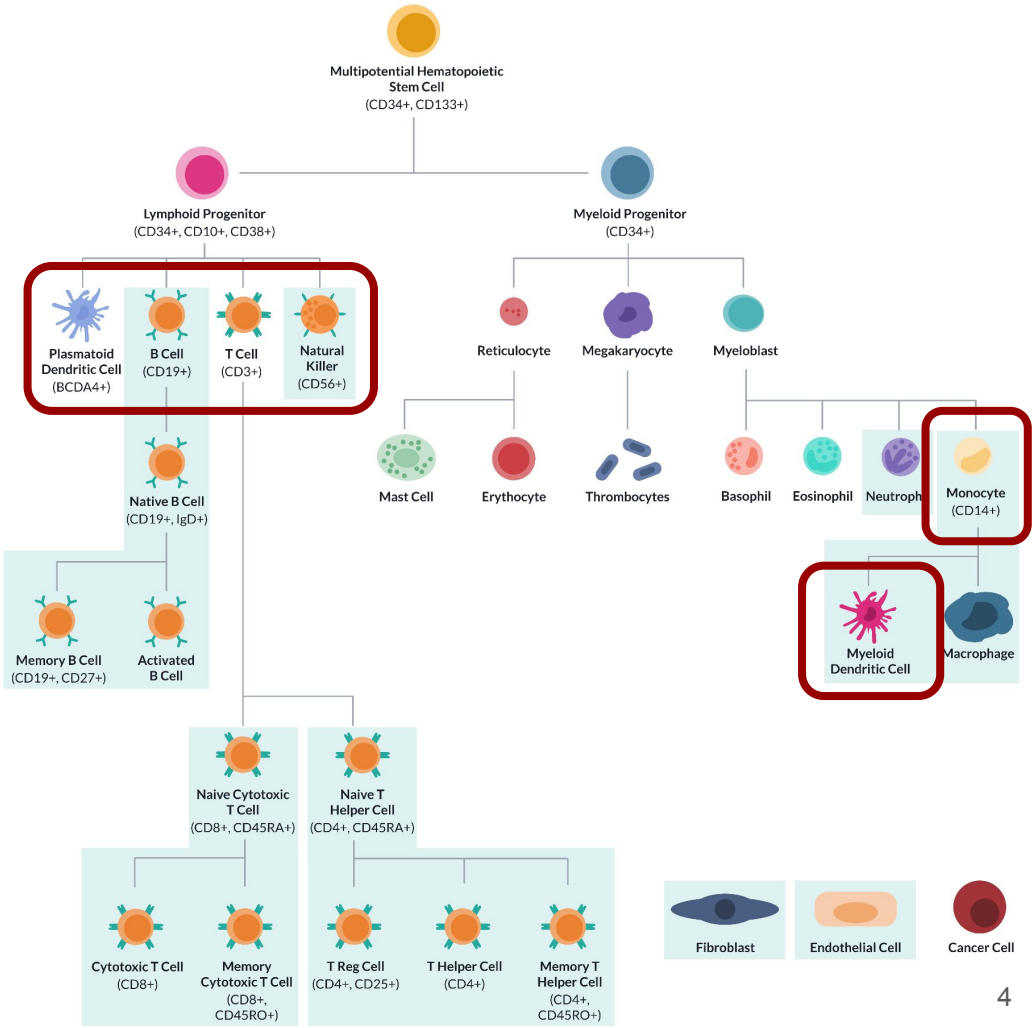
Signature matrix creation is an important step in accurate bulk RNA-seq deconvolution







Aim: Accurate deconvolution of immune cell types from tumor bulk RNA-seq



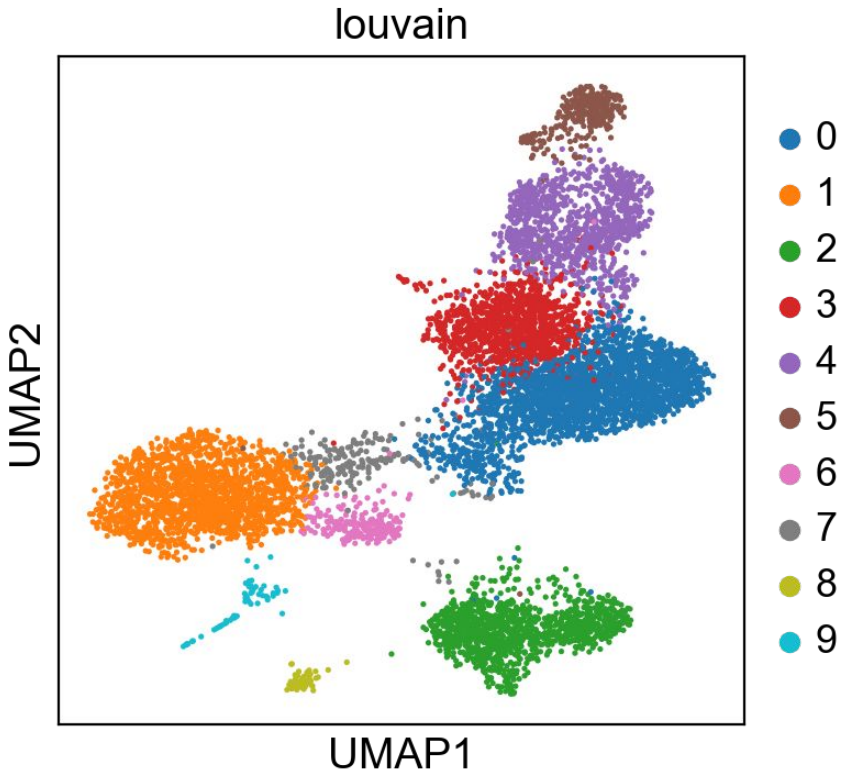
Methods: Baseline signature matrix

- Starting from the LM22 matrix (Newman, A.,et al., 2015)
  - 22 leukocyte subsets profiled on the HGU133A platform
  - Significantly differentially expressed genes between each populations to rest using a two-sided unequal variance t-test
  - Ranked by fold change and selected gene features with the lowest condition number
  - Additional GSEA and highly expressed genes in non-hematopoietic cancer cell lines
- Adopted the LM22 signature matrix → selecting highest expressed proportion of the 5 cell types we are testing
  - Tried running Cibersort on nearly full LM22 and merging (poorer results)

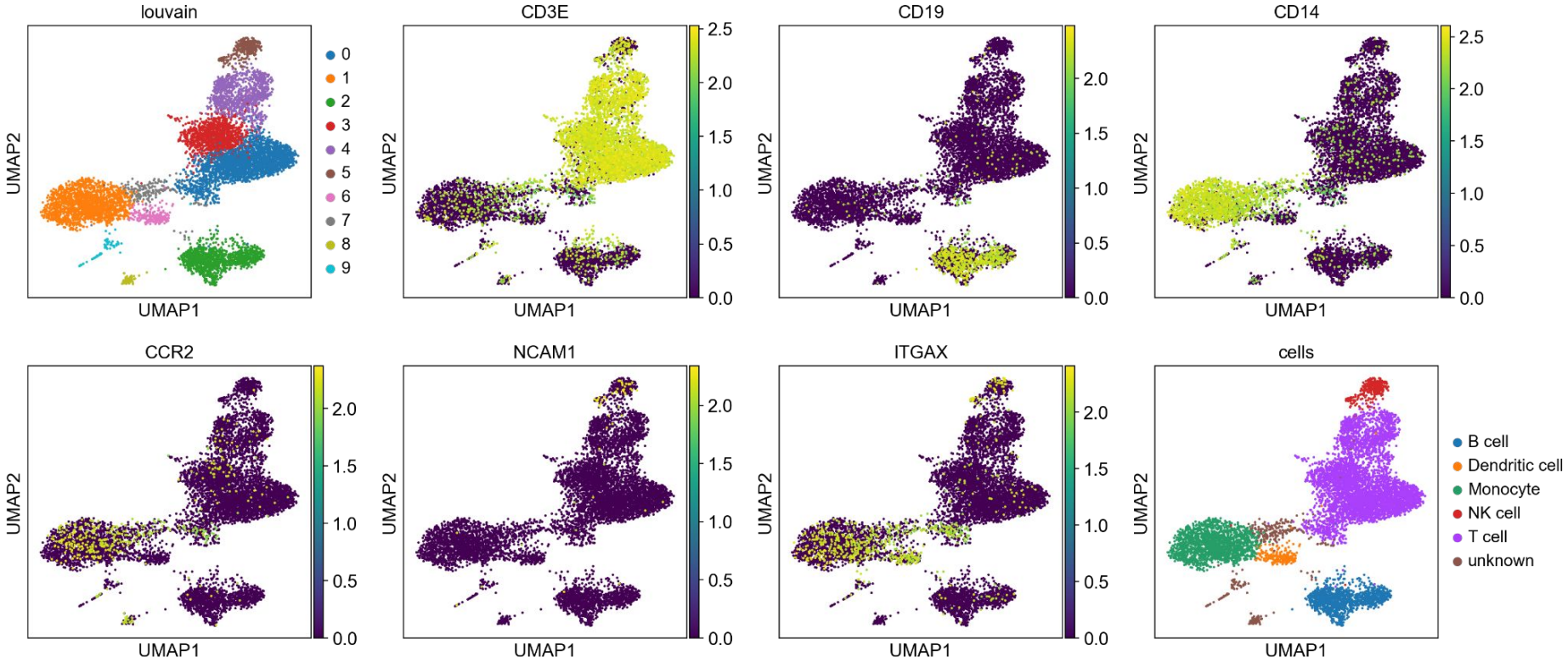
Signature matrix	**PBMC_example_signature_matrix_v1.tsv**	**LM5.txt**
Correlation (pearson)	0.34821	0.52271
RMSE	0.055	0.105

Methods: Dataset selection and pre-processing

- Dataset selection: (cells x genes)
  - PBMC_6kv1_tpm.csv: 5419 x 32738
  - **PBMC_8kv2_tpm.csv: 8381 x 33694**
    - Largest dataset; w/o combining due to potential batch effect
  - PBMC_10kv3_tpm.csv: 7865 x 33555
- Standard scRNA-seq pre-processing:
  - Filter genes expressed in less than 3 cells
  - Filter cells:
    - >10% mt content
    - <200 genes
    - >6000 genes
  - Identified 3365 highly variable genes
- Clustering:
  - Louvain clustering (resolution = 1.0)
  - Final dimensions: 8381 x 3365 (cells x genes)

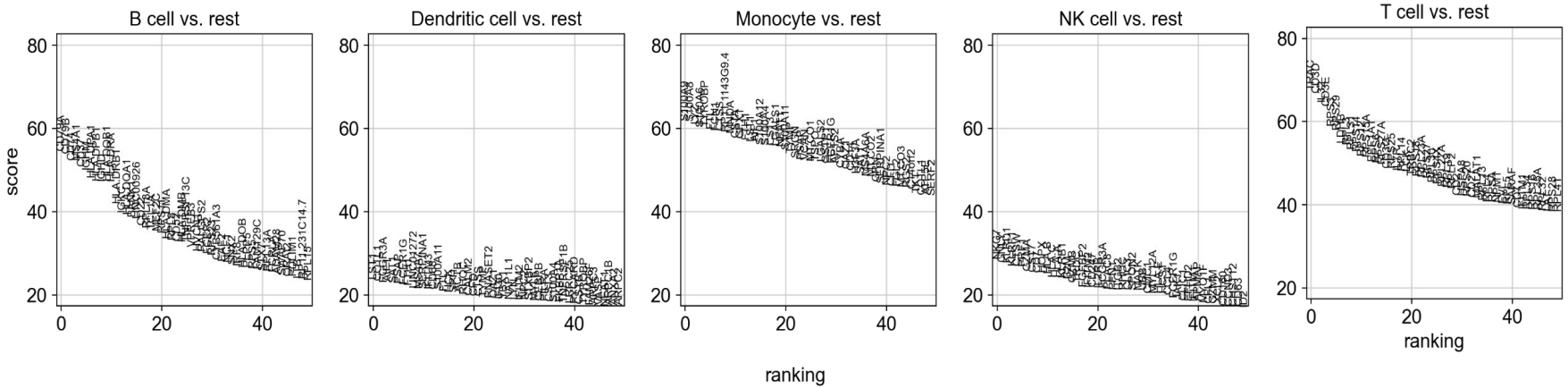


Methods: Cell type annotation by gene markers



Methods: Creation of signature matrix

- Differential gene expression using scanpy (sc.tl.rank_genes_groups)
  - Cell cluster vs. rest
  - Wilcoxon signed-rank test (non-parametric test)
  - Benjamini-Hochberg p-value correction
- Filter for features by p-value and fold change
- **Signature matrix:** Matrix of raw expression of selected genes for centroid of cell types clusters

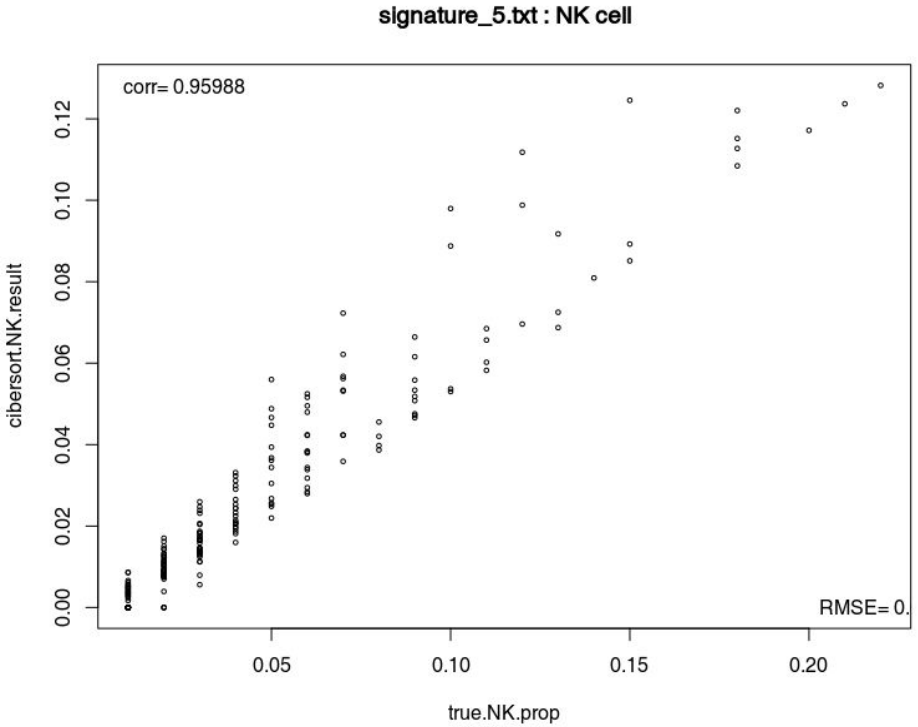


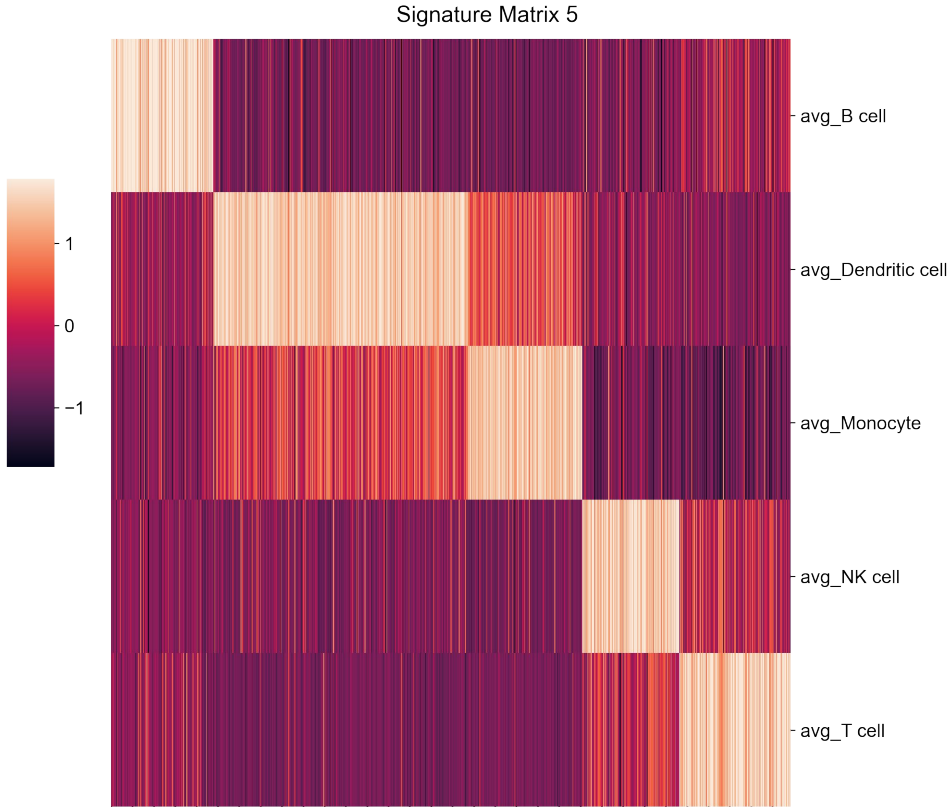
Results: Variations on signature matrix creation

- Benchmarked on: CellLineA_Tumor_bulk_composition_noise_0.tsv
- Signatures 1-3: vary across DC cluster annotation
- Signatures 4-7: vary p-value and fold change thresholds
- Metrics across vectorized matrices:
  - Pearson correlation
  - RMSE

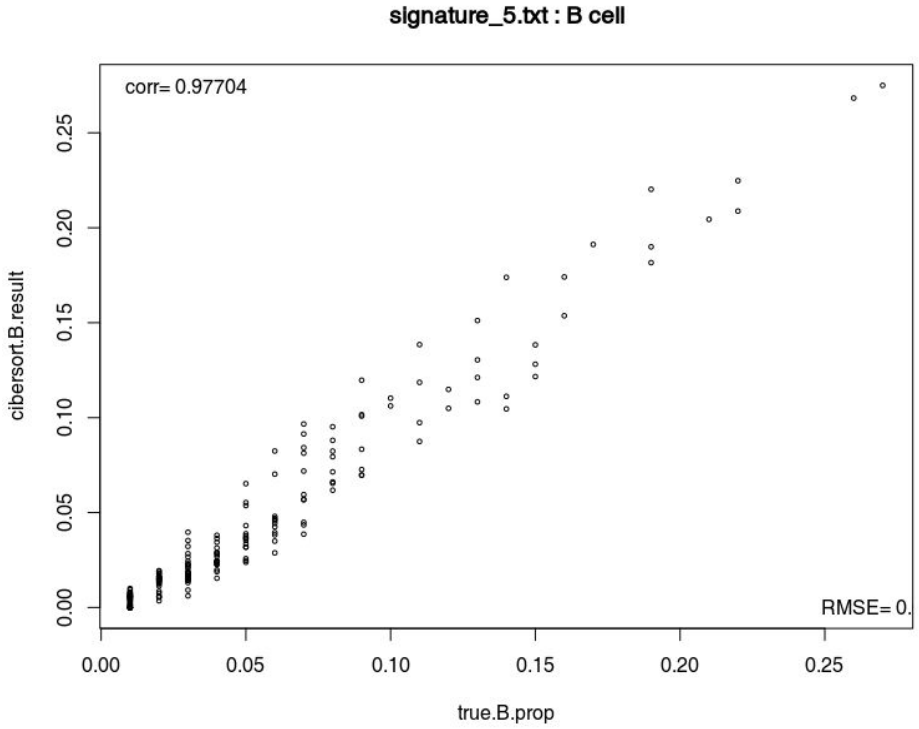
Sig.	eg.	LM5	my_LM5	sig_1	sig_2	sig_3	sig_4	sig_5	sig_6	sig_7
Cor	0.34821	0.52271	0.77079	0.68186	0.84013	0.61636	0.79812	0.83402	0.79455	0.83630
RMSE	0.055	0.105	0.052	0.044	0.030	0.041	0.032	0.026	0.032	0.030
n_genes	665	548	456	744	792	704	521	989	466	1387
padj	na	na	na	0.05	0.05	0.05	0.1	1e-8	0.025	0.3
fc	na	na	na	3	3	3	3.5	2	3.5	2.5

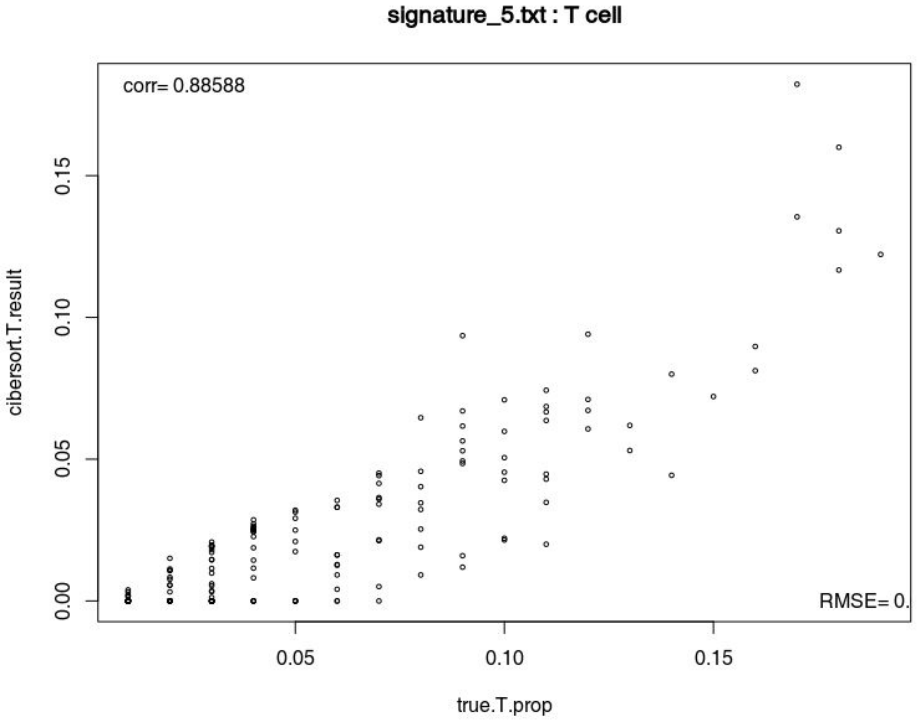
Results: Signature Matrix 5 (NK cell, heatmap)



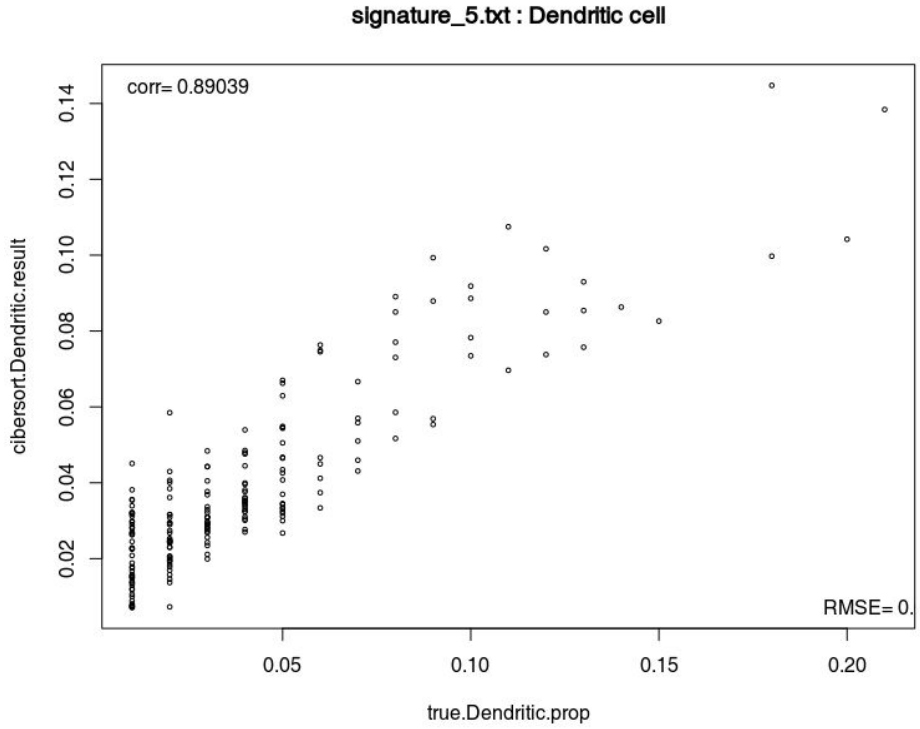


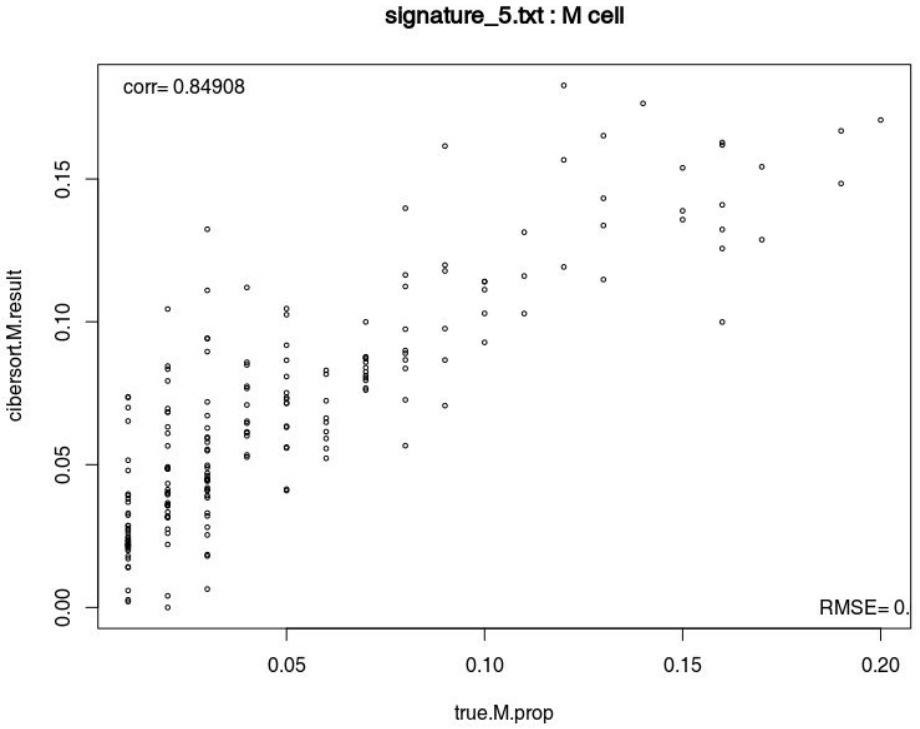
Results: Signature Matrix 5 (B cell, T cell)





Results: Signature Matrix 5 (Dendritic cell, M cell)





Future directions

- Generate signature matrices across additional datasets
- Integrate/adopt methodology and datasets mentioned in the creation of LM22 in Cibersort
  - Filtering by GSEA and highly expressed genes in non-hematopoietic cancer cell lines
- Additional searching for optimal p-value and fold change thresholds
  - Measure scalability of a larger signature matrix
  - Additional visualizations of signature matrix and features search
- Add additional cell types (semi-fine grain)