Subject Section

NITUMID: Nonnegative Matrix Factorization-based Immune-TUmor MIcroenvironment Deconvolution

Daiwei Tang [1,a], Seyoung Park [2,a] and Hongyu Zhao [1,*]

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT, 06511, USA and
[2]Department of Statistics, Sungkyunkwan University, Jongno-gu, Seoul, South Korea

*To whom correspondence should be addressed.
a: These authors contributed equally to this work.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: A number of computational methods have been proposed recently to profile tumor microenvironment from bulk RNA data, and they have proved useful for understanding microenvironment differences among therapeutic response groups. However, these methods are not able to account for tumor proportion nor variable mRNA levels across cell types.
Results: In this article, we propose a Non-negative Matrix Factorization-based Immune-TUmor MIcroenvironment Deconvolution (NITUMID) framework for tumor microenvironment profiling that addresses these limitations. It is designed to provide robust estimates of tumor and immune cells proportions simultaneously, while accommodating mRNA level differences across cell types. Through comprehensive simulations and real data analyses, we demonstrate that NITUMID not only can accurately estimate tumor fractions and cell types' mRNA levels, which are currently unavailable in other methods; it also outperforms most existing deconvolution methods in regular cell type profiling accuracy. Moreover, we show that NITUMID can more effectively detect clinical and prognostic signals from gene expression profiles in tumor than other methods.
Availability: The algorithm is implemented in R. The source code can be downloaded at https://github.com/tdw1221/NITUMID.
Contact: hongyu.zhao@yale.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

During tumor development, tumor-host immune interaction shapes a complex tumor microenvironment (TME) consisting of tumor cells, various infiltrating immune cells, and other non-cancerous cells. TME is both clinically and prognostically informative (Zhang *et al.*, 2003; Bindea *et al.*, 2013; Angelova *et al.*, 2015).

With the success of checkpoint blockade immunotherapy, there is a clear need for accurate characterization of TME because the effectiveness of checkpoint blockade requires the presence of infiltrating immune cells (Ribas and Wolchok, 2018). It has been found that the TME profiles might be associated with response and clinical benefits (Riaz *et al.*, 2017a). Conventional methods such as flow-cytometry and immuno-histochemistry (IHC) are not practical for massive profiling due to their high cost, labor, and tissue availability. An alternative approach to this problem is through *in silico* analysis of the abundant available bulk microarray and RNA-Seq data from tumor tissues since these data also contain valuable genomics information of all cells in these (admixed) samples.

Accurate estimation of the proportions of different cell types from these data poses a number of challenges, including determining important candidate cell types in the complex microenvironment (Hanahan and Weinberg, 2011), defining effective signature genes for each cell type given much similarity among some of the immune cell types, and accounting for technical and biological variation of measured gene expression levels, particularly distinct total mRNA amounts among different cell types (Racle *et al.*, 2017).

A number of methods have been proposed to infer cell type proportions from bulk expression data in the literature. Single-sample Gene Set

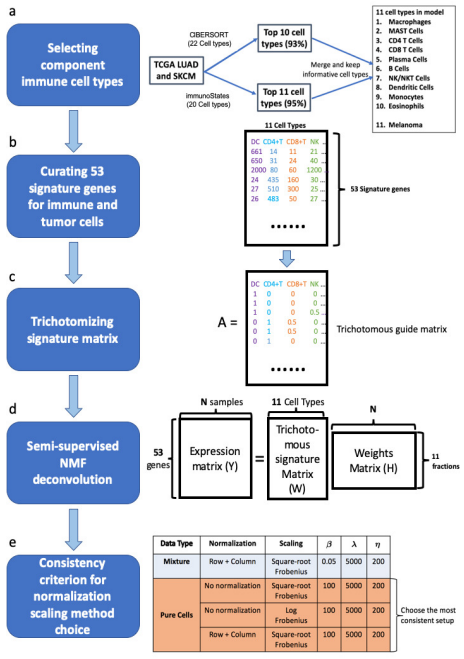
Enrichment Analysis (ssGSEA) based methods, such as ESTIMATE (Yoshihara *et al.*, 2013) and xCell (Aran *et al.*, 2017), leverage cell-type specific gene set information to estimate the abundances of cell types. However, score-based methods are not able to estimate comparable proportions across cell types. Another class of methods are regression-based (Newman *et al.*, 2015; Li *et al.*, 2016; Schelker *et al.*, 2017; Racle *et al.*, 2017; Vallania *et al.*, 2018), which are usually based on a self-curated signature matrix consisting of cell-type specific signature genes and their representative expression levels. It was found that such fixed-signature based methods are sensitive to signatures and not able to fully account for the biological and technical variations (Vallania *et al.*, 2018). There are also two additional issues that have not received enough attention: 1) Modeling and inference of tumor proportion, which has remained challenging due to significant heterogeneity of cancer expression; and 2) accounting for different mRNA levels across cell types.

To address the limitations in the existing methods, we present a Nonnegative Matrix Factorization-based Immune-TUmor MIcroenvironment Deconvolution (NITUMID) framework in this paper that features a flexible Nonnegative Matrix Factorization (NMF) deconvolution procedure and a trichotomous signature matrix. NITUMID adaptively infers signature matrix and cell proportions in a semi-supervised fashion. Compared with fixed signature matrix, our framework is more flexible for accounting for technical and biological variances, thus could obtain more robust cell type proportion estimation. NITUMID also includes tumor as a component cell type, which enables us to infer the proportions of tumor and immune cells simultaneously, instead of only relative proportions of immune cell types. Due to the availability of single cell tumor RNA-seq data during NITUMID's development, the results presented below focus on melanoma microenvironment as proof-of-concept.

To ensure the computational stability of NITUMID, our framework is based on the uniqueness of factorization (see Proposition 1). We use data driven method based on a consistency criterion to choose tuning parameters, including "optimal" data normalizations and scalings for each run of NITUMID. Due to input matrix Y's structural difference between pure cell and cell mixture case, we use different sets of preprocessing methods/parameters in these two cases (Figure 1e).

We evaluated the performance of NITUMID by applying it to gene expression data collected from purified cells, synthetic cell mixtures, and bulk melanoma data sets. We found that NITUMID outperformed most existing methods on microarray, bulk, and single cell RNA-Seq gene expression data. More importantly, we also demonstrated that adjustment of mRNA differences and simultaneous estimating tumor and immune cells fractions, can lead to biologically more informative results compared to the existing methods.

Fig. 1: Diagram of NITUMID. **a.** We applied CIBERSOT (Newman *et al.*, 2015) and immunoStates (Vallania *et al.*, 2018) on the TCGA SKCM and LUAD data and selected the intersection of the most abundant cell types from these two methods as our component cell types. **b.** We curated a list of 53 signature genes for the 11 cell types and obtained their mean expression profiles in each cell type. **c.** We trichotomized the signature genes expression profile matrix into matrix A. **d.** Illustration of the NMF framework in NITUMID: for the 53 by N gene expression matrix Y, we input the guide matrix A, and factorize matrix Y into a 53 by 11 matrix W and an 11 by N matrix H. **e.** We designed and implemented a consistency-based criterion to choose the model's tuning parameters for different datasets. See Supplementary section A.5 for details.



2 Methods

2.1 Problem Setup

We formulate the deconvolution problem through Non-negative Matrix Factorization (NMF) (Figure 1d). Given an observed gene expression matrix $Y = [Y_1, \cdots, Y_n] \in R^{m \times n}$, where $Y_i \in R^m$ represents the i-th sample's expression profile consisting of m genes, our goal is to decompose Y into the product of two matrices: $W \in R^{m \times r}$ and $H \in R^{r \times n}$, where m, n, and r are the numbers of genes, samples, and cell types (subtypes), respectively:

$$Y = \Big[Y_1, \ \cdots \ , Y_n\Big] = \Big[W_1, \ \cdots \ , W_r\Big] \cdot \Big[H_1, \ \cdots \ , H_n\Big] = WH,$$

where $W_k \in R^m$ can be interpreted as a representative expression profile of a specific subtype k, $H_i \in R^r$ represents the proportions of component cell types (subtypes) in sample i, thus they sum up to 1, $W \in R_+^{m \times r}$, and $H \in R_+^{r \times n}$. Our method, as described below, is essentially a variation of NMF (Lee and Seung, 2000).

The baseline framework for the proposed decomposition can be expressed as the following optimization problem:

$$\min_{W \in R_+^{m \times r},\ H \in R_+^{r \times n}} \|Y - WH\|_F^2 \ s.t. \ \sum_{k=1}^{r} H_{ki} = 1 \text{ for all } i. \quad (1)$$

Note that the constraint $\sum_{k=1}^{r} H_{ki} = 1$ is essential to have meaningful solutions for H, and the above problem becomes conventional NMF (Lee and Seung, 2000) without that constraint. However, there is no guarantee that solving (1) reveals true biological signals encoded in Y, especially when data Y is corrupted or contains problematic (noisy) genes. Moreover, the NMF results may not be unique, which can lead to numerical instability. To address these problems, we instead consider a set of signature genes that are informative predictors of their corresponding cell types, which can be selected based on prior biological knowledge of each cell type (See Supplementary Section A.1). Suppose that we have m_k signature genes for the k-th cell type, with a total of $m = \sum_{k=1}^{r} m_k$ signature genes across cell types. The ideal case is when each signature gene is only expressed in one cell type, where the corresponding W should have orthogonal columns as $W_1, \cdots, W_r$ have mutually exclusive support set. However, signature genes of a specific cell type are often expressed in other cell types as well, e.g. *NAPSB* is usually highly expressed in the *Dendritic cell*, but also

has moderate expression values in the *Macrophage*, *Monocyte*, and *B cell* types. Selecting mutually exclusive signature genes across different cell types may also be unrealistic between biologically similar cell types such as *CD4* and *Regulatory T cell*. Here to address this issue, we assign different weights to our signature genes based on their capacity in differentiating cell types.

More specifically, for each cell type k, we define primary marker gene sets G_k^{prime} and secondary marker gene sets G_k^{second}, where G_k^{prime} contains signature genes that are only highly expressed in cell type k, while G_k^{second} consists of genes that are substantially expressed in cell type k but other cell types also have intermediate expression levels of those genes, thus genes in G_k^{second} can be a primary marker gene of other cell types. For simplicity, we use G_k^{p} and G_k^{s} to denote G_k^{prime} and G_k^{second}, respectively. For example, *NAPSB* belongs to $G^{\text{p}}_{Dendritic}$, while it is also a component of $G^{\text{s}}_{Macrophage}$, $G^{\text{s}}_{Monocyte}$, and G^{s}_{Bcell}. For simplicity, let $G_k := G_k^{\text{p}} \cup G_k^{\text{s}}$ be the set of marker genes for cell type $k = 1, \cdots, r$. The NITUMID framework is based on the following optimization problem by incorporating the signature gene information into NMF:

$$\min_{W \in S_W,\, H \in S_H} \frac{1}{2}\|Y - WH\|_F^2, \tag{2}$$

where S_W and S_H are sets of W and H, respectively:

$$S_W = \{W \in \mathbb{R}_+^{m \times r} \mid \text{Supp}(W_{\cdot,k}) = G_k,\ \min_{j \in G_k^{\text{p}}} W_{jk} \geq \max_{j \in G_k^{\text{s}}} W_{jk}\}$$

$$S_H = \{H \in \mathbb{R}_+^{r \times n} \mid 1_r^T H = 1_n^T\}.$$

2.2 NITUMID

Implementation of the above deconvolution framework to melanoma datasets involves three major steps: determining the component cell types, finding signature genes for each cell type, and choosing appropriate parameters. The component cell types in the model should be both abundant in TME while having biological importance. Since melanoma and lung adenocarcinoma are two major targets of immuno-oncology, we began by profiling their immune cell composition with two existing deconvolution methods: CIBERSORT (Newman *et al.*, 2015) and immunoStates (Vallania *et al.*, 2018). By merging the major immune cell types identified by both methods, we ended up with a list of 10 immune cell types, and melanoma is the 11-th cell type in the model (Figure 1a, See Supplementary Section A.2 for more details).

As described above, finding signature genes that are exclusively-expressed in certain cell type would be ideal but challenging. So we followed a proceudure described in Supplementary Section A.3 to find exclusive markers first and include some secondary markers where exclusive markers are not available. We ended up with 53 signature genes for these 11 cell types and translated the signature matrix into a trichotomous guide matrix $A \in \mathbb{R}^{m \times r}$ (Figures 1b-c, Supplementary_Table1, Supplementary Section A.3), where each entry has a value of 1, 0.5, and 0, respectively. For a specific cell type, an entry 1 means that the corresponding gene is highly expressed in that cell type, an entry 0 means that the gene is not expressed, and an entry 0.5 means that the gene has an intermediate expression level in that cell type. To more effectively account for differences of the primary and secondary signature genes, we incorporate the following weight matrix $\tilde{A} = \mathbb{E} - A$ in the NMF framework, where $\mathbb{E}$ is an m by r matrix with all entries 1.

Finally, we propose the following NMF, namely, "NITUMID", by incorporating the signature gene information using regularization terms:

$$\min_{W \geq 0, H \geq 0} \frac{1}{2}\|Y - WH\|_F^2 + \lambda \|\tilde{A} \odot W\|_1 + \frac{\beta}{2}\|W\|_F^2 + \frac{\eta}{2}\|1_r^T H - 1_n\|^2, \tag{3}$$

where Y is a gene expression matrix, and $\odot$ indicates element-wise product, and λ and η are regularization parameters that balance the approximation error $\|Y - WH\|_F^2$ and the structure constraints for W and H, respectively. The rational behind $\|\tilde{A} \odot W\|_1$ is that by penalizing the non-zero entries in W using the weight as in $\tilde{A}$, we expect that W_{jk} is penalized less when the j-th gene is a signature gene of the k-th cell-type. The matrix $\tilde{A}$ can be regarded as a semi-signature matrix as it contains some information of signature genes, but does not specify their expression levels. The idea is to maintain the relative mRNA level structure while granting enough flexibility to "learn" mRNA levels from the observed data.

This type of weighted ℓ_1 penalty is a common way to enforce a specific structure of the target matrix or vector. It is worth noting that the second penalty term $\frac{\beta}{2}\|W\|_F^2$ is added to make (3) strictly convex with respect to W to improve the convergence of the algorithm as shown in Proposition S4. This quadratic (smoothness) constraint is often enforced to regularize the solutions in the presence of noise in the data (Berry *et al.*, 2007). Note that without this second penalty term, (3) is not necessarily strictly convex with respect to W during the iterative algorithm shown later, unless the columns of H are full-rank during iterations. However, the term $\|H\|_F^2$ is not necessary because (3) is already strictly convex with respect to H due to the term $\frac{\eta}{2}\|1_r^T H - 1_n\|^2$. Note that larger η yields faster convergence of the algorithm (See Proposition S4).

Finally, to ensure the stability and robustness of NMF across different input Y, we considered four different combinations of scaling, normalization, and parameter choices. Specifically, one set of these procedures and parameters is for bulk gene expression data while the rest three are for pure cell gene expression data (Figure 1e). For the pure cell data case, results from the most consistent specification is chosen among the three candidates. The details of the consistency criterion and choices of regularization parameters are given in Supplementary Sections A.4 and A.5, respectively. If the type of input data matrix is not known, we suggest using the mixture mode in general. We have demonstrated that for pure cell data matrix Y, the mixture mode shows comparable results with other deconvolution methods although the pure cell mode generally outperforms all (See Section 3.1 for details). Only when it is known that the input data are from pure cells, it is preferred to specify NITUMID in the pure cell mode.

2.3 Uniqueness property

In this subsection, we prove the uniqueness of the proposed NMF decomposition when the NMF factors satisfy certain constraints. The uniqueness of NMF factors is essential when interpreting the estimated proportions. Without uniqueness, one can not assert that the obtained NMF factors reveal true biological information even when the underlying true factors are in the underlying spaces, i.e. $W^o \in S_W$ and $H^o \in S_H$. Throughout this section, we assume $r \leq n$.

The following Proposition 1 shows that if each cell type has a primary marker gene, then the uniqueness of the proposed NMF is guaranteed. All the proofs are deferred to Supplementary Sections C and D.

Proposition 1. *Let $Y = W^o H^o$ be the underlying decompositions of Y satisfying $W^o \in S_W$ and $H^o \in S_H$. Suppose that H^o is a full rank matrix. Suppose for each cell type k, there exists a gene j_k such that $j_k \in G_k \setminus (\cup_{l \neq k} G_l)$. Then, there exists a unique decomposition of Y in the sense that $Y = WH$ with $W \in S_W$ and $H \in S_H$ implies $W = W^o$ and $H = H^o$.*

In the Supplementary Section D, we prove that the proposed NMF enjoys unique decomposition property under more general conditions such that some cell types do not need to have exclusive marker genes (See Proposition S6). Although in our real data, all cell types have their own exclusive marker genes, we investigate the usefulness of the proposed NMF when such cell type exists through simulation. We observe that NITUMID

still effectively infers cell types for this case. See Supplementary Section B.2 for details.

2.4 Algorithm

The proposed NMF optimization problem (3) is solved by alternatively updating W and H using the following multiplicative update rules:

Input: an m by n nonnegative matrix Y.
Output: an m by r nonnegative matrix W and a r by n nonnegative matrix H.
Step 1: At $t = 0$, set the initial nonnegative matrices $W^{(0)} = E_{m\times r} - A$ and $H^{(0)} = E_{r\times n}$, where $E_{m\times r}$ is the m by r matrix with all ones.
Step 2: At the t-th iteration, we update $W^{(t)}$ and $H^{(t)}$

$$W_{jk}^{(t)} = \frac{W_{jk}^{(t-1)}(Y(H^{(t-1)})')_{jk}}{(W^{(t-1)}H^{(t-1)}(H^{(t-1)})')_{jk} + \lambda A_{jk} + \beta W_{jk}^{(t-1)} + 10^{-16}},$$

$$H_{ki}^{(t)} = \frac{H_{ki}^{(t-1)}\{((W^{(t)})'Y)_{ki} + \eta\}}{((W^{(t)})'W^{(t)}H^{(t-1)})_{ki} + \eta 1_r' H_i^{(t-1)} + 10^{-16}}$$

until the stopping criterion is satisfied.

Note that 10^{-16} is added to denominators to avoid division by zero. And we use the same $W^{(0)}$ and $H^{(0)}$ to avoid randomness in the obtained outcome. The computational complexity of NITUMID is $O(mnr)$, same as that of NMF. Our iterative algorithm guarantees the convergence to the local optimum of (3) as shown in Proposition 2. The monotonic convergence of the algorithm can be proven using an auxiliary function approach similar to that used to prove the convergence of the Expectation-Maximization algorithm. Note that since the proposed optimization is not convex and we utilize a multiplicative update algorithm, the best computational convergence result one could obtain is convergence to some stationary point (Lin, 2007; Nakano *et al.*, 2010) as we show in Proposition 2. Proposition 2 also shows that any iterate $W^{(t)}$ has an ideal structure in the sense that it has zero value for any entries (j, k) when gene j is not a signature gene for cell type k.

Proposition 2. *Let $E(\cdot,\cdot)$ be the objective function in* (3)*. Then the iterates $\{W^t\}_{t\geq 0}$ and $\{H^t\}_{t\geq 0}$ of the proposed algorithm converges to a stationary point of* (3) *with the following rate:*

$$\min_{t\in[1,K]} \beta\|W^{t+1}-W^t\|^2+\eta\|H^{t+1}-H^t\|^2 \leq \frac{2}{(K-1)}E(W^0,H^0).$$

Moreover, for $j \notin G_k$, it holds that $W_{jk}^{(t)} = 0$ for any t.

Proof. The proof is deferred to Supplementary Sections C and D. We use the auxiliary function approach, following Zhang *et al.* (2008), Ding *et al.* (2010), and Shang *et al.* (2016).

3 Results

3.1 Performance on Pure Cell Data

We first benchmarked NITUMID's performance on pure cell data (both microarray and scRNA-Seq) along with eight other deconvolution methods. As mentioned earlier in Section 2.2 and Figure 1e, NITUMID has both the mixture mode and the pure cell mode. Since the data nature is not always known in practice, we benchmarked both modes on our testing data. The results are shown in Supplementary Section B.1.

Across the 19 microarray and scRNA-Seq datasets, the NITUMID mixture mode had comparable results with other deconvolution methods. When it is specified that the data come from pure cells, the NITUMID pure mode could consistently outperform other deconvolution methods (Supplementary Section B.1 and Figure S3ab). However, we do acknowledge that this true information is only used in the NITUMID pure mode, whereas the other methods do not use this information.

3.2 Performance on *in silico* synthetic mixture data

In this section, we investigated and benchmarked NITUMID's performance via *in silico* mixture data. We began with evaluating how NITUMID's performance changes with respect to sample size. Each *in silico* cell mixture sample was generated as follow. For each component cell type, we randomly chose one sample from our training microarray data matrix of that cell type, and this would form a ten-column matrix; then we generated a random Dirichlet sample of length 10, and right multiplied the matrix to get a convex combination of ten sample columns (See Supplementary Section A.8 for details). We considered a sample size $n \in \{50, 200, 500, 1000\}$, and investigated NITUMID's performances through 50 runs. Figure S1a shows the boxplots of the 50 median values of the correlations from n samples. It can be seen that the proposed NMF generally gave accurate predictions of H across different values of n, and larger n led to better accuracy.

Next, we compared NITUMID's performance with four major existing methods (immunoStates, CIBERSORT, xCell, and EPIC) on *in silico* mixture. To account for the imbalanced mRNA across cell types, we considered two types of *in silico* mixture: unweighted and weighted mixture. The unweighted mixture was described above. For weighted mixture, each cell sample was first scaled by a cell type-specific factor before mixing. The factors are rough estimates from Figure 3a (DC=1, CD4=1, CD8=1, Macrophage=10, Bcell=4, NK/NKT=6, Monocyte=6, Plasma=10, MAST=1, esinopil=1). Here on top of our training microarray datasets, we also collected scRNA-Seq immune samples from two other studies as pool for *in silico* mixture generation (TME contains 3,208 cell samples of T cell, B cell, NK cell, Macrophages, and dendritic cell (Tirosh *et al.*, 2016); 10X has 596 samples of CD4+ T cell, CD8+ T cell, NK cell, B cell and monocytes (Zheng *et al.*, 2017)). The simulation results (sample size 50) are shown in Figures 2a-c.

Here we can see that when weights were introduced, all methods had reduced accuracy. In microarray training data (Figure 2a), NITUMID had good performance as expected. For the TME case, although NITUMID, CIBERSORT, and xCell had similar performance in the unweighted scenario, NITUMID maintained better accuracy when weighted (Figure 2b). NITUMID also maintained a small advantage for the 10X case (Figure 2c). Although *in silico* mixture might not fully characterize real bulk data, these simulation results suggest that NITUMID's semi-supervised setup does have an advantage when mRNA levels are imbalanced.

We then conducted model-based simulations where the nonzero entries of the underlying signature matrix W were generated from a normal distribution while each column of H follows the Dirichlet distribution. For this experiment, we considered Model 1 and Model 2. (See Supplementary Section A.9 for details). Figure S1b shows the distribution of samples' median correlations from 50 runs across the different noise levels.

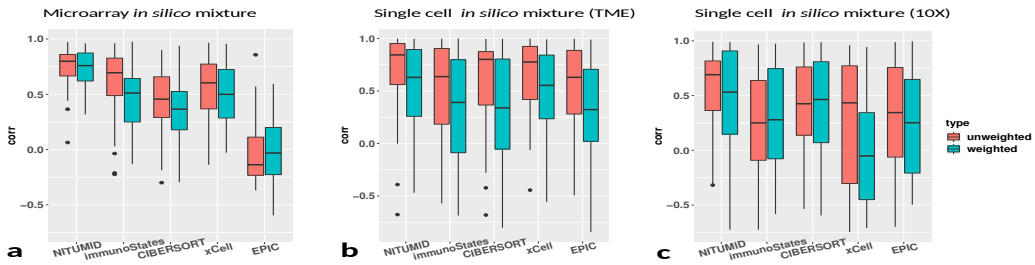
It can be seen that NMF generally had better accuracy for Model 1 since Model 2 has more cell types, which is generally more challenging.

We also examined the stability of the inferred H in terms of randomly selected subsamples from the entire dataset, where we would like to see stable estimates across different subsets of samples. See Supplementary Section B.7 for details of the experiment and result. It can be seen that the estimated cell proportions of NMF are highly robust with respect to the random subsets of samples for all the cases (Correlation > 0.98), suggesting the stability of the proposed NMF method.

3.3 NITUMID is able to capture biologically informative signals in real bulk tumor data

In this section, we applied our method to four bulk melanoma datasets: (1) the NIVO cohort, consisting of 118 melanoma RNA-Seq samples

Fig. 2: Comparison of NITUMID, immunoStates, CIBERSORT, xCell and EPIC's performances on *in silico* immune cell mixture, measured by Pearson correlation between estimated cell fractions and true cell fractions. **a** *in silico* immune cell mixture is generated from our microarray training datasets; **b** *in silico* immune cell mixture is generated from melanoma tumor microenvironment (TME) scRNA-Seq data (Tirosh *et al.*, 2016) ; **c** *in silico* immune cell mixture is generated from 10X scRNA-Seq immune cell profiles from (Zheng *et al.*, 2017)



from 65 patients in an anti-PD-1 Nivolumab therapy trial (56 pre-treatment and 62 on-treatment) (Riaz *et al.*, 2017a); (2) the HUGO cohort, consisting of 27 pre-treatment metastatic melanoma RNA-Seq samples of anti-PD-1 therapy (Hugo *et al.*, 2016); (3) the CTLA4 cohort, consisting of 42 pre CTLA4-treatment tumor RNA-Seq samples (Allen *et al.*, 2015); and (4) the TCGA skin cutaneous melanoma (SKCM) cohort, downloaded from the MD Anderson TCGA Batch effect tool (https://bioinformatics.mdanderson.org/main/TCGABatchEffects:Overview), where batch effects were removed with the empirical Bayes method. It consists of 368 immunotherapy-free SKCM RNA-Seq samples. We used RPKM as gene expression measures for all four datasets.

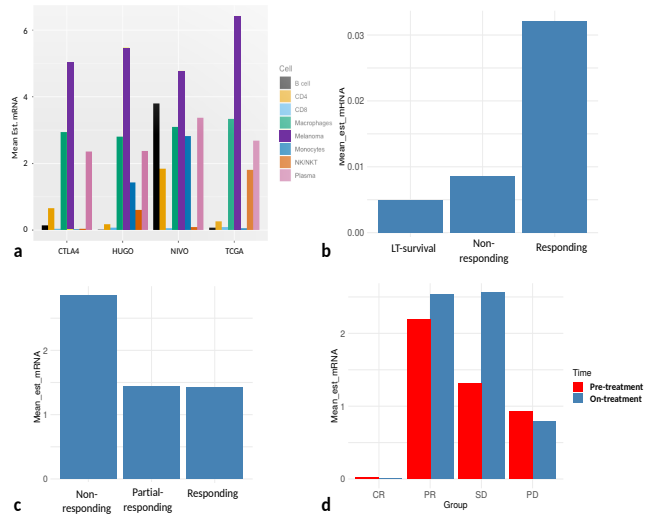
3.3.1 NITUMID can reconstruct cell type-specific mRNA level structure

One advantage of not using fixed signature matrix W in NITUMID is that it can adaptively learn the signature genes' expression levels in a data-dependent manner. This is particularly important for tumor-immune deconvolution since the component cell types tend to have different total mRNA contents.

Figure 3a shows the average signature gene mRNA levels by cell types from the obtained $\widehat{W}$ matrix. It can be seen that the inferred mRNA levels have a relatively consistent pattern across datasets, with high mRNA levels of melanoma cells, plasma cells, and macrophages/monocytes. This is largely consistent with our prior biological knowledge since plasma and melanoma are usually in active metabolism, while macrophages/monocytes are larger in size. Racle *et al.* (2017) showed that macrophages/monocytes' total mRNA is much higher than that of NK, B, and T cells, which can also be observed in our learned $\widehat{W}$. We also observe that B cells, NK/NKT cells, and monocytes' mRNA levels showed some variations, which is probably due to the fact that the transcription levels in these cells can be significantly affected by the cell's status (*e.g.* B cells become plasma cells, monocytes become macrophages) or the spillover effects between them.

The results of mRNA estimation are also consistent with clinical findings. Allen *et al.* (2015) found that samples from the responding group showed higher cytolytic activity, and we can see the same trend from Figure 3b, where CD8+ T cell's mRNA level is elevated in long term survival and responding group. From our estimated mRNA levels from the HUGO cohort (Hugo *et al.*, 2016) (Figure 3c) and the NIVO cohort (Riaz *et al.*, 2017a) (Figure 3d), we can see that the complete responding (CR) group consistently has lower macrophage mRNA levels, which is in line with findings from two recent studies (Lavin *et al.*, 2017; Cassetta and Kitamura, 2018). These results suggest that NITUMID's estimated mRNA level can have clear biological interpretation and implication.

Fig. 3: Estimated mean mRNA levels **a**. Estimated mRNA levels by cell type from the W matrix of the four bulk melanoma datasets. **b**. CD8+ T cells mRNA levels for the CTLA4 dataset, by response group. **c**.Macrophages mRNA levels for the HUGO dataset by response group **d**.Macrophages mRNA levels for the NIVO dataset by the response group and treatment status, partial responding (PR), stable disease (SD) and progressive disease (PD) groups all showed higher levels compared with the complete responding (CR) group

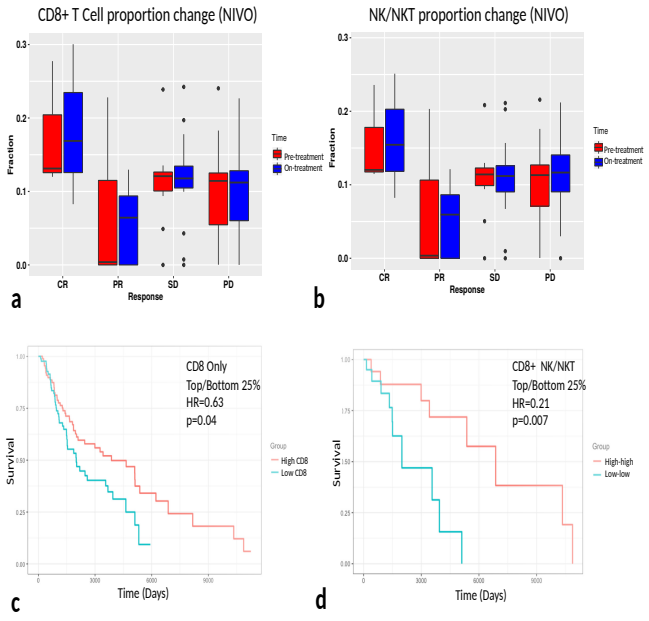


3.3.2 NITUMID detects tumor proportion change in the immunotherapy responding group

Another advantage of NITUMID is simultaneous estimating tumor and immune cells fractions. Here we present results of NITUMID-estimated

tumor fraction from bulk melanoma. The estimated tumor fraction of the pre-treatment and on-treatment NIVO cohort samples showed substantial reduction in the complete responding (CR) group (p=0.10), while the reductions of the other groups are less significant (Figure S15a). This suggests that NITUMID is capable of detecting the tumor fraction change. For the pre-treatment data from the CTLA4 and HUGO cohorts (Figures S15b-c), the tumor fractions across different groups are relatively similar. The similarity in fractions is expected since pre-treatment patients are not expected to differ substantially in tumor proportions.

Fig. 4: Informative immune cell fractions changes identified by NITUMID **a**. Estimated CD8+ T cell fraction by response group and treatment status (NIVO). **b**. Estimated NK/NKT cell fraction by response group and treatment status (NIVO). **c**. Survival between high CD8+ T cell and low CD8+ T cell groups in the TCGA SKCM cohort, the high group includes 92 samples from the first CD8+ fraction quantile (top 25%), the low group includes the fourth quantile (bottom 25%). **d**. Survival difference between samples with high CD8+ T and high NK/NKT cell proportions and those with low proportions in both (TCGA SKCM). High-high group consisting of patients whose CD8+ T and NK/NKT cell fractions are both in top 25%, the low-low group consisting of patients whose CD8+ T and NK/NKT cell fractions are both in the bottom 25%



3.3.3 NITUMID reveals immune cells fraction change during therapy and identifies informative treatment response predictors

Previous research has established association between anti-PD-1 immunotherapy response and elevated CD8+ T cell and NK/NKT cell levels (Tumeh *et al.*, 2014; Iraolagoitia *et al.*, 2016; Liu *et al.*, 2017; Eroglu *et al.*, 2018). Here, we applied NITUMID to the NIVO dataset, where patients were treated with anti-PD-1 drug Nivolumab (Riaz *et al.*, 2017b). Figures 4a and 4b show that this pattern of higher CD8+ T cells and NK/NKT cells fractions in the complete responding group (CR) can be revealed by NITUMID. Interestingly, the results show higher fractions regardless of the treatment status for both cell types. However, this pattern is not observed in CD4+ T cells (Figure 5e), while macrophages (Figure 5d) show a reversed pattern where the pre-treatment complete responding group has lower macrophages fraction, which is consistent with our previous mRNA results.

The above results suggest that higher fractions of CD8+ T cells and NK/NKT cells might be associated with better survival regardless of treatment status. To further test this pattern, we stratified samples from the TCGA SKCM cohort (367 samples) by their estimated CD8+ T and NK/NKT fractions. Since TCGA samples had distinct treatment statuses and were all immunotherapy free, these patterns should preserve. For CD8+ T cells only, samples with higher fractions had significant survival benefits compared with those with lower fractions (Figures 4c and 5a, Log-rank test p-values 0.04 and 0.03); but NK/NKT fraction itself does not give significant results (Figure 5b, Log-rank test p-value 0.87). When combined together, patients with higher fractions in both groups still had survival advantages over those with low fractions in both (Figures 4d and 5c, Log-rank test p-values 0.007 and 0.1). These results support the patterns identified previously and suggest that estimates from NITUMID might be prognostically informative.

We also applied CIBERSOT, xCell, and EPIC (Figures S16, S17, and S18) to these data sets. These methods have inconsistent patterns on the NIVO cohort in terms of CD8+ T cells and NK/NKT cells, and none of them shows the same concordant pre-/on-treatment higher CD8 and NK/NKT pattern as NITUMID (Figures S16ab, S17ab, and S18ab). For the TCGA SKCM dataset, samples stratified by the estimated CD8+ T cell fraction showed significant differences in survival for all four methods, while only EPIC and NITUMID's NK/NKT fraction estimates can stratify groups with significant survival differences when combined with CD8+ T cell fractions (Figures 4cd, S16cd, S17cd, and S18cd).

We also investigated the association between macrophages' fraction and survival, by stratifying the TCGA SKCM cohort into high (top 50%) and low (bottom 50%) macrophage groups (Figure 5f). We obtained a 1.1 hazard ratio for the high macrophage group, which is consistent with our previous findings, but the result is not statistically significant (p-value=0.50).

Allen *et al.* (2015) demonstrated that in the CTLA-4 blockade therapy, the responding group and long-term survival group have higher cytolytic activity level than that in the non-responding group, and the long term survival group is even higher than that in the responding group (Allen *et al.*, 2015). We applied NITUMID to their data and also observed the same trend in the CD8+ T cell fractions (Figure 5g), although the difference is not statistically significant. Similar results are given by CIBERSORT, NITUMID, and xCell (Figures S19a-c).

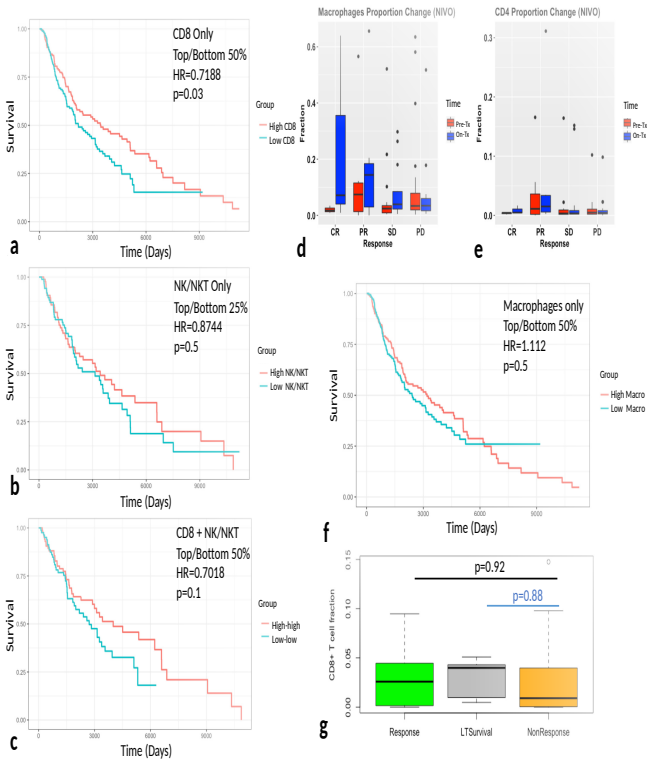
These results suggest that on top of having the advantage of being able to estimate tumor fractions, NITUMID also shows potential to detect clinically informative signals in immune cell fractions.

4 Discussion

We have proposed an NMF-based algorithm NITUMID for tumor-immune microenvironment deconvolution. NITUMID features a semi-supervised flexible framework that not only takes into account different mRNA levels across cell types, but also provides proportion estimates for both tumor and immune cells.

The baseline framework for NITUMID guarantees unique decomposition with pre-specified structure of the signature genes matrix W and column sum condition for H, overcoming the identifiability issues faced by classical NMF framework. As for data normalization and scaling, NITUMID considers several parameter specifications and chooses the optimal one with the most robust results. NITUMID also enjoys fast convergence that has been empirically observed in various data sets (See Supplementary Section C).

Fig. 5: Additional immune cell fractions-related results **a**. Survival curves for the high and low CD8+ T cell groups for TCGA SKCM stratified by the median CD8+ T cell fraction (184 samples each group). **b**. Survival curves for the high and low NK/NKT cell groups for TCGA SKCM stratified by the median NK/NKT fraction (184 samples each group). **c**. TCGA SKCM survival stratified by estimated CD8+ T cell and NK/NKT fractions. The high-high group are 19 samples whose both fractions are in top 50%, while the low-low group are 19 samples whose both fractions are in bottom 50%. **d-e**. Estimated Macrophage and CD4+ T cell fractions by response group and treatment status for the NIVO. **f**. Survival curves for the high and low macrophages groups for TCGA SKCM stratified by the median macrophages fraction. **g**. Estimated CD8+ T cell fraction by response group for the CTLA-4 data (Allen *et al.*, 2015)



NITUMID involves penalty parameters λ, β, and η which are determined by the Karush-Kuhn-Tucker condition combined with consistency test with respect to random permutation and additive errors in the data (See Supplementary Sections A.4 and A.5). This novel parameter tuning strategy may also be applied to other methods involving regularization parameters.

By benchmarking NITUMID onto multiple gene expression datasets of purified cells, bulk tumor, as well as *in silico* cell mixtures, we demonstrated that NITUMID can more accurately infer the proportions of different immune and tumor cells. We also showed that NITUMID's semi-supervised set-up of "learning" cell type-specific mRNA levels can produce biologically and clinically reasonable values. Finally, we demonstrated that the immune and tumor cells proportions inferred from NITUMID can be informative on patients' prognosis and treatment response.

One potential issue with NITUMID is that its usage of the semi-supervised guide signature matrix A might result in information loss compared with exact signature gene expression matrix. We evaluated this by generating simulated data using the exact W from CIBERSORT to compare the performances of CIBERSORT (with the exact gene expression matrix) and NITUMID (with the trichotomous guide matrix). The results showed that with increasing noise level, NITUMID gradually outperformed exact regression (Figure S8), which suggests that under real scenario in which both biological and technical variances exist, NITUMID can still maintain an edge in balancing accuracy and robustness. See Supplementary Section B.5 for more details. NITUMID is also robust to data format, normalization, and batch effect removal methods (Supplementary Section B.4), as well as signature genes number (Supplementary Section B.8), which enables it to be widely applicable.

Current version of NITUMID only works with melanoma-immune cell microenvironment. However, several potential issues may affect NITUMID's future generalization into other cases, including component immune cell types, model consistency when adding/removing cell types, among others. Some exploratory analysis has been done to consider those issues: by exploring other major tumor types' microenvironment, we found that the 10 immune cell types in NITUMID account for $>99\%$ of immune cells (Supplementary Section B.6.1). We also showed the estimated cell fractions are largely consistent when component cell types change (Supplementary Section B.6.2). Finally, we offered preliminary results for two extension of NITUMID that we applied to breast cancer and upper respiratory microenvironment, respectively (Supplementary Section B.6.3). With those results together, we believe that with the accumulation of more cancer genomics data, NITUMID's framework can be generalized to accommodate more cell types as well as complex microenvironments.

Acknowledgements

In the development process of NITUMID, Dr. Victor Du and Dr. Jungmin Choi from Yale School of Medicine provided us with valuable insights into tumor microenvironment and its compositions. Professor Katerina Politi and Steven Kleinstein from Yale School of Medicine have offered us insightful comments on this project, both biologically and computationally. We also thank Dr. Ying Zhu and Dr. Yiyi Liu for helpful discussions.

Funding

Daiwei Tang and Hongyu Zhao were partially funded by NIH grants R01GM122078 and 3P50 CA196530. Seyoung Park was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2019R1C1C1003805).

References

Allen, E. M. V. *et al.* (2015). Genomic correlates of response to ctla4 blockade in metastatic melanoma. *Science*, pages 207–211.

Angelova, M. *et al.* (2015). Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.*, **16**, 1–17.

Aran, D., Hu, Z., and Butte, A. J. (2017). xcell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.*, **18**, 1–14.

Berry, M. *et al.* (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, **52**, 155–173.

Bindea, G. *et al.* (2013). Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*, **39**, 782–795.

Cassetta, L. and Kitamura, T. (2018). Targeting tumor-associated macrophages as a potential strategy to enhance the response to immune checkpoint inhibitors. *Frontiers in cell and developmental biology*, **6**, 38.

Ding, C., Li, T., and Jordan, M. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(1), 45–55.

Eroglu, Z., Zaretsky, J. M., Hu-Lieskovan, S., Kim, D. W., Algazi, A., Johnson, D. B., Liniker, E., Kong, B., Munhoz, R., Rapisuwon, S., *et al.* (2018). High response rate to pd-1 blockade in desmoplastic melanomas. *Nature*.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, **144**, 646–674.

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., *et al.* (2016). Genomic and transcriptomic features of response to anti-pd-1 therapy in metastatic melanoma. *Cell*, **165**(1), 35–44.

Iraolagoitia, X. L. R., Spallanzani, R. G., Torres, N. I., Araya, R. E., Ziblat, A., Domaica, C. I., Sierra, J. M., Nuñez, S. Y., Secchiari, F., Gajewski, T. F., *et al.* (2016). Nk cells restrain spontaneous antitumor cd8+ t cell priming through pd-1/pd-l1 interactions with dendritic cells. *The Journal of Immunology*, **197**(3), 953–961.

Lavin, Y., Kobayashi, S., Leader, A., Amir, E.-a. D., Elefant, N., Bigenwald, C., Remark, R., Sweeney, R., Becker, C. D., Levine, J. H., *et al.* (2017). Innate immune landscape in early lung adenocarcinoma by paired single-cell analyses. *Cell*, **169**(4), 750–765.

Lee, D. D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, **13**, 556–562.

Li, B. *et al.* (2016). Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biol*, **17**, 1–16.

Lin, C. J. (2007). On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, **18**, 1589–1596.

Liu, Y., Cheng, Y., Xu, Y., Wang, Z., Du, X., Li, C., Peng, J., Gao, L., Liang, X., and Ma, C. (2017). Increased expression of programmed cell death protein 1 on nk cells inhibits nk-cell-mediated anti-tumor function and indicates poor prognosis in digestive cancers. *Oncogene*, **36**(44), 6143.

Nakano, M. *et al.* (2010). Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with Î²-divergence. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*.

Newman, A. M. *et al.* (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, **12**(5), 453–457.

Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E., and Gfeller, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife*, **6**.

Riaz, N. *et al.* (2017a). Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell*, **171**(4), 934–949.

Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., Hodi, F. S., Martín-Algarra, S., Mandal, R., Sharfman, W. H., *et al.* (2017b). Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell*, **171**(4), 934–949.

Ribas, A. and Wolchok, J. D. (2018). Cancer immunotherapy using checkpoint blockade. *Science*, **359**(6382), 1350–1355.

Schelker, M. *et al.* (2017). Estimation of immune cell content in tumour tissue using single-cell rna-seq data. *Nat. Commun.*, **8**(2032).

Shang, R., Zhang, Z., Jiao, L., Wang, W., and Yang, S. (2016). Global discriminative-based nonnegative spectral clustering. *Pattern Recognition*, **55**, 172–182.

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., *et al.* (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, **352**(6282), 189–196.

Tumeh, P. C., Harview, C. L., Yearley, J. H., Shintaku, I. P., Taylor, E. J., Robert, L., Chmielowski, B., Spasic, M., Henry, G., Ciobanu, V., *et al.* (2014). Pd-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature*, **515**(7528), 568.

Vallania, F., Tam, A., Lofgren, S., Schaffert, S., Azad, T. D., Bongen, E., Haynes, W., Alsup, M., Alonso, M., Davis, M., *et al.* (2018). Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nature communications*, **9**(1), 4735.

Yoshihara, K. *et al.* (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, **4**(2612).

Zhang, J. *et al.* (2008). Pattern expression nonnegative matrix factorization: Algorithm and applications to blind source separation. *Computational Intelligence and Neuroscience*, pages 1–10.

Zhang, L. *et al.* (2003). Intratumoral t cells, recurrence, and survival in epithelial ovarian cancer. *N Engl J Med*, **348**, 203–213.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., *et al.* (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, **8**, 14049.