HHS Public Access

Author manuscript

Cytometry A. Author manuscript; available in PMC 2022 September 01.

Published in final edited form as:

Cytometry A. 2021 September ; 99(9): 899–909. doi:10.1002/cyto.a.24298.

Ab initio spillover compensation in mass cytometry data

Qi Miao[1,2], **Fang Wang**[1], **Jinzhuang Dou**[1], **Ramiz Iqbal**[1], **Muharrem Muftuoglu**[3], **Rafet Basar**[3], **Li Li**[3], **Katy Rezvani**[3], **Ken Chen**[1]

[1]Department of Bioinformatics and Computational Biology, Division of Quantitative Sciences, The University of Texas MD Anderson Cancer Center, Houston, Texas

[2]Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas

[3]Department of Stem Cell Transplantation and Cellular Therapy, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas

Abstract

Signal intensity measured in a mass cytometry (CyTOF) channel can often be affected by the neighboring channels due to technological limitations. Such signal artifacts are known as spillover effects and can substantially limit the accuracy of cell population clustering. Current approaches reduce these effects by using additional beads for normalization purposes known as single-stained controls. While effective in compensating for spillover effects, incorporating single-stained controls can be costly and require customized panel design. This is especially evident when executing large-scale immune profiling studies. We present a novel statistical method, named CytoSpill that independently quantifies and compensates the spillover effects in CyTOF data without requiring the use of single-stained controls. Our method utilizes knowledge-guided modeling and statistical techniques, such as finite mixture modeling and sequential quadratic programming, to achieve optimal error correction. We evaluated our method using five publicly available CyTOF datasets obtained from human peripheral blood mononuclear cells (PBMCs), C57BL/6J mouse bone marrow, healthy human bone marrow, chronic lymphocytic leukemia patient, and healthy human cord blood samples. In the PBMCs with known ground truth, our method achieved comparable results to experiments that incorporated single-stained controls. In datasets without ground-truth, our method not only reduced spillover on likely affected markers, but also led to the discovery of potentially novel subpopulations expressing functionally meaningful, cluster-specific markers. CytoSpill (developed in R) will greatly enhance the execution of large-scale cellular profiling of tumor immune microenvironment, development of

Correspondence Ken Chen, Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77230.

AUTHOR CONTRIBUTIONS

Qi Miao: Conceptualization; data curation; formal analysis; methodology; software; visualization; writing-original draft; writing-review and editing. **Fang Wang:** Methodology; writing-review and editing. **Jinzhuang Dou:** Methodology; writing-review and editing. **Ramiz Iqbal:** Writing-review and editing. **Muharrem Muftuoglu:** Conceptualization; data curation; resources. **Rafet Basar:** Conceptualization; data curation; resources. **Li Li:** Conceptualization. **Katy Rezvani:** Supervision. **Ken Chen:** Conceptualization; funding acquisition; methodology; project administration; resources; supervision; writing-review and editing.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

novel immunotherapy, and the discovery of immune-specific biomarkers. The implementation of our method can be found at https://github.com/KChen-lab/CytoSpill.git.

Keywords

compensation; CyTOF; mass cytometry; spillover; statistical methods

1 | INTRODUCTION

The emergence and rapid adoption of mass cytometry (CyTOF) as a more scalable alternative to flow cytometry has led to unprecedented fine-grained profiling of human cell populations. CyTOF employs metal-isotope-tagged monoclonal antibodies to measure the expressions of the surface proteomic markers and/or intracellular signaling molecules in single cells. This technology is particularly important for various biomedical fields, such as immunology, oncology, and stem cell research, because CyTOF allows experimentalist to measure between 40 and 60 parameters (channels) from around 100,000 cells in a single assay. CyTOF has been widely used to profile the immune system in the presence of disease such as the characterization of cellular heterogeneity in tumor samples [1, 2]. The advantage of this technology over its predecessor is the minimal amount of spectral overlap between channels that is more typical in flow cytometry and can lead to differing interpretations of cell populations [3, 4]. Despite this advantage, spillover effects similar to those in the flow cytometry data is still observed, since the intensity measured in a channel can be affected by the intensity of the neighboring channels.

Spillover effects, while generally minor, can substantially limit the accuracy of cell type identification. It is possible to alleviate this by selecting high purity isotopes, redesigning metal isotopes, or using control panel, but that approach is complicated, time-consuming, and costly [5]. These approaches can be even more burdensome in the context of executing a large-scale immune profiling study where spillover effects might not be easily compensated for. Given that the spillovers have an approximately linear relationship with respect to the original signal [5], error reduction can be achieved by transforming the data using a properly estimated spillover matrix without necessitating any changes to study design or materials.

Here, we present a novel computational method, CytoSpill, that can independently compensate the spillover effects in CyTOF data without using any single-stained controls. Our method utilizes knowledge-guided modeling and statistical algorithms to infer the optimal spillover matrix and perform correction. We utilize the knowledge about spillover sources to constrain the estimation of the spillover matrix. The underlying assumption of our method is that the spillover component can be separated from the signal component in affected channels using a mixture distribution model.

2 | MATERIALS AND METHODS

2.1 | Datasets

We examined our method using five CyTOF datasets, obtained from peripheral blood mononuclear cells (PBMCs) with a 36-antibody panel [5], C57BL/6J mouse bone marrow

with a 38-antibody panel [6], healthy human bone marrow with a 32-antibody panel [7], chronic lymphocytic leukemia patient blood with a 46-antibody panel and healthy human cord blood sample with a 44-antibody panel. Table 1 presented the details of the datasets we used. All the datasets used were deposited in flow repository with ID: FR-FCM-Z2KW and on https://github.com/KChen-lab/CytoSpill.

2.2 | Spillover compensation problem

In CyTOF data, spillover usually comes from three sources: abundance sensitivity, isotopic impurity and oxidization [8]. The information on these three sources can be obtained based on the isotopes used in the experiment panel. If an isotope used in one channel has a similar mass to the isotope used in another channel, for example, differed by 1 atomic mass, these channels will affect each other (abundance sensitivity). Channels using isotopes with the same metal will affect each other (isotopic impurity). If the atomic mass of the isotope in a channel is larger than that in another channel by 16, the atomic mass of an oxygen atom, it will potentially be affected by this other channel through oxidation, since the metal may get oxidized incidentally during the experiment. In CyTOF experiments, cells expressing a marker will have positive readings while cells that do not express a marker will have zero readings. However, if a channel is affected by spillover, these cells may acquire some level of intensity readings, resulting from spillover effects. The signal artifacts resulting from spillover effects are usually smaller than the true biological signals, thus adding a low background modal to the marker expression density. We assume that the density of a marker affected by spillover follows a multimodal distribution, where the lowest intensity component corresponds to the spillover noise and the other components correspond to true expression levels.

The spillover compensation problem in CyTOF data is similar to the complete compensation problem defined in flow cytometry literate [9]. Let

$$\boldsymbol{D}_{N\times M} = \boldsymbol{T}_{N\times M}\boldsymbol{S}_{M\times M},$$

where $\boldsymbol{D}_{N\times M}$ is the matrix of observed signals, $\boldsymbol{T}_{N\times M}$ is the matrix of true signals and $\boldsymbol{S}_{M\times M}$ is the spillover matrix whose diagonal elements are all 1's. N is the number of cells and M is the number of channels (refer to Table 2 for commonly used notations in this article). To model the spillover noise, we define the noise components as $\boldsymbol{Y}_{N\times M}$ that Thus, we have

$$\boldsymbol{Y}_{N\times M} = \boldsymbol{D}_{N\times M} - \boldsymbol{T}_{N\times M}.$$

Thus, we have

$$\boldsymbol{Y} = \boldsymbol{T}(\boldsymbol{S} - \boldsymbol{I}).$$

Since all diagonal elements of $\boldsymbol{S}$ are 1's, $(\boldsymbol{S} - \boldsymbol{I})$ is a matrix contains only the off diagonal elements of $\boldsymbol{S}$ with $\boldsymbol{I}$ being an identity matrix. If we can estimate $\boldsymbol{S}$, we can perform compensation via $\boldsymbol{T} = \boldsymbol{D}\boldsymbol{S}^{-1}$. $\boldsymbol{S}$ can be usually estimated by employing single-stained

controls [5], which can be time-consuming and costly. Here, our goal is to derive ***S*** through modeling of ***Y***. The mathematical problem is that only ***D*** is observed and that ***Y*** and ***T***, both having high dimensions, will need to be determined simultaneously from their relationship to ***D***. We hypothesize that it is possible to achieve this by making appropriate assumptions on ***Y*** and constraining the structure as well as the parameters in ***S*** using prior knowledge about the spillover structures. Based on our assumption that the noise component ***Y*** in each channel forms a lower modal in that channel's intensity density distribution, we can model channel intensities using mixture probability distributions and segregate the noise modals from the mixed signals.

2.3 | Cutoff derivation

In order to separate the noise component from the true signal component, we derived a cutoff value for each channel in each CyTOF dataset. We assume that the intensity observed in each channel follows a mixture of Gaussian distributions where the Gaussian with the lowest mean represents the spillover noise. We fit the intensity distribution in a channel j (j = 1, 2, … M) by a finite mixture model using function initFlexmix from R package FlexMix, assuming there are K (K = 1, 2…5) components [10]. The cells with zero intensity were excluded from the model. It does not have a principal limit on the number of components. We set a limit of 5 for limiting the computation cost for this procedure. The observed marker expression level in channel j follows a multi-normal distribution:

$$d_j \sim \sum_{k=1}^{K} \alpha_{kj} F_{kj}$$

We choose the model with the lowest integrated completed likelihood value. If the model suggests more than one component (K > 1), we will derive a cutoff value c assuming a type 1 error a = 0.05. K = 2 suggests that the channel has a bimodal distribution and the lower modal is the noise component. The cutoff value c is then derived as the probability of c belonging to the lower modal:

$$\mathrm{P}(c \in F_1) = \frac{\alpha_1 F_1(c)}{\alpha_1 F_1(c) + \alpha_2 F_2(c)} = 1 - a = 0.95.$$

If K > 2, it suggests that the channel has a multimodal distribution. In that case, we will take the lowest modal and the highest modal to calculate c, assuming that the highest modal corresponds to the true signal. If the returned model only has one component, we will select an empirical cutoff at 10% quantile of the channel intensity. This is an important step in our method which defined the error components contributed by spillover effects on negative cells of the markers. These error components will be further utilized in next steps for estimating the spillover coefficients. Given the derived cutoff values $\boldsymbol{C} = \{c_1, c_2, c_3 \ldots c_M\}$, the noise component ***Y*** is defined as

$$Y_{i,j} = \begin{cases} D_{i,j} & \text{if } D_{ij} < c_j \\ 0 & \text{otherwise} \end{cases}.$$

2.4 | Estimation of the spillover matrix

S quantifies the effect of spillover between the channels in the panel. The off-diagonal values represent the fraction of the signal in channel A that gets added into channel B. We attempted two methods to estimate the spillover matrix $\boldsymbol{S}$**:** (1) sequential quadratic programming (SQP) with channel-specific constraints and (2) non-negative matrix factorization (NMF). The prior knowledge of error sources, that is, abundance sensitivity, isotopic impurity and oxidization are utilized as constraints for model optimization.

2.4.1 | Method 1: SQP with channel-specific constraints—We define a channel interaction matrix $\boldsymbol{Z}_{M\times M}$ from the prior knowledge described above. $\boldsymbol{Z}$ consists of the off-diagonal elements of $\boldsymbol{S}$ that $\boldsymbol{S} = \boldsymbol{Z} + \boldsymbol{I}$. The external information passed in via Z matrix reflects an intrinsic property of the CyTOF assay, (e.g., which channels [antibodies] are tagged by isotopes of similar mass). These physical–chemical properties lead to spillover in three sources: abundance sensitivity, isotopic impurity and oxidization. For a new dataset, the structure of Z is determined based on the design of the assay (i.e., the physical–chemical properties of the isotopes and the technology). To facilitate this, we implemented a function in our R package that recognizes the isotope information that the users provide in the CyTOF data and generates a Z matrix that reflects the expected spillover patterns. $Z_{i,j}$ is greater than 0 only when channel i can potentially affect channel j and 0 otherwise. Because spillover has strict additive effect [5], $Z_{i,j}$ is not allowed to have negative values. Furthermore, previous studies indicated that the spillover effects are generally less than 10% [5]. Thus, we apply a boundary constraint: $Z_{i,j} \leq 0.1$ to the estimation. We estimate the channel interaction matrix using a SQP algorithm [11] as follows:

$$\min_{\boldsymbol{Z}} \| \boldsymbol{Y} - \boldsymbol{D}\boldsymbol{Z} \|^2_F ,$$

where $\boldsymbol{Z}$ satisfies $0 \leq Z_{i,j} \leq 0.1$ when channel i can potentially affect channel j and $Z_{i,j} = 0$ otherwise. Here, $\|\cdot\|^2_{\mathrm{F}}$ denotes the Frobenius norm of a matrix. The spillover matrix $\boldsymbol{S}$ is calculated as $\boldsymbol{S} = \boldsymbol{Z} + \boldsymbol{I}$.

2.4.2 | Method 2: NMF—Since $\boldsymbol{D} = \boldsymbol{T}\boldsymbol{S}$ where $\boldsymbol{T}$ and $\boldsymbol{S}$ are both non-negative by definition, we can also formulate the task of spillover matrix estimation as a masked NMF problem [12–14]. We incorporate the cutoffs we derived and the prior knowledge into the NMF model using binary masks. With the derived cutoff values $\boldsymbol{C}$, we generate a binary matrix $\boldsymbol{B}_{N\times M}$ that masks non-noise component defined by the cutoff value in $\boldsymbol{D}$ that

$$B_{i,j} = \begin{cases} 1 \text{ if } D_{ij} < c_j \\ 0 \text{ otherwise} \end{cases}.$$

Given the channel interaction matrix $\boldsymbol{Z}$ defined above, we have:

$$\boldsymbol{D} = \boldsymbol{T}\boldsymbol{S},$$

$$\boldsymbol{D} = \boldsymbol{T}(\boldsymbol{Z} + \boldsymbol{I}),$$

$$\boldsymbol{B} \cdot \boldsymbol{D} = \boldsymbol{B} \cdot (\boldsymbol{T}(\boldsymbol{Z} + \boldsymbol{I})).$$

The noise mask $\boldsymbol{B}$ ensures that only the error component selected by our derived cutoffs can affect the optimization. We assuming the error component is normally distributed under our assumption. The resulting method performs NMF by solving the following optimization problem:

$$\min_{\mathbf{T} \geq 0, 0 \leq \boldsymbol{Z} \leq 0.1} \| \boldsymbol{B} \cdot \boldsymbol{D} - \boldsymbol{B} \cdot (\boldsymbol{T}(\boldsymbol{Z} + \boldsymbol{I})) \|^2_F .$$

The optimization is performed using gradient decent algorithm in Tensorflow [15].

2.5 | Compensation

Given the estimated spillover matrix $\boldsymbol{S}$, we can obtain the real (compensated) data $\boldsymbol{T}$ via the following optimizing problem:

$$\min_{\boldsymbol{T}} \| \boldsymbol{D} - \boldsymbol{T}\boldsymbol{S} \|^2_F \quad s.t.\ \boldsymbol{T} \geq 0,$$

using a non-negative least squares algorithm [5, 16]. The $\boldsymbol{T} \geq 0$ since CyTOF data does not contain negative values.

2.6 | Evaluation

To evaluate our method, we compared the compensated data using the spillover matrix generated by our method with the compensated data using the single-stained controls. We also compared the uncompensated data with compensated data using four other datasets that do not have single-stained controls. We examined the data clustering results using t-SNE plots generated with Rtsne package followed by PhenoGraph clustering [7, 17, 18].

2.7 | Simulation

We first used simulation to test the ability of SQP and NMF to estimate the spillover matrix in our model. We simulated true signals $\boldsymbol{T}$ with 20 channels and 100,000 cells. We simulated 80 cell populations each with 1250 cells in the data. For each population, there are 15 channels with positive expressions and 5 channels with 0 expressions as negative channels. For each positive channel in each population, the expression values followed a normal distribution with mean randomly drawing from a uniform distribution U(100, 600) and SD equal to mean/5.

Based on our assumptions on error sources, we obtained the structure of the simulated spillover matrix. In each of the simulated spillover matrix $\boldsymbol{S}_{M \times M}$ where $M = 20$, there are 105 off diagonal elements of spillover coefficients that need to be simulated. Moreover,

the spillover coefficients were simulated based on a uniform distribution U(0, 0.1). We first simulated 20 spillover matrices. We then applied these spillover matrices to the simulated true signal $\boldsymbol{T}$ to generate the simulated CyTOF data $\boldsymbol{D}_{\text{simu}}$ with spillover effects in it: $\boldsymbol{D}_{\text{simu}} = \boldsymbol{T}\boldsymbol{S}_{\text{simu}}$. We then performed estimation on $\boldsymbol{S}_{\text{simu}}$ using our approaches and compared the estimated spillover matrices with the known simulated spillover matrices.

3 | RESULTS

In order to compensate CyTOF data without using single-stained controls, we made assumptions that spillover noise contributes as a new modal at the lower end of the affected channel expression density. We derived a cutoff for each channel of the data to separate the noise modal from observed signal modal. We then assumed that the spillover effects come from three main sources which will constrain the spillover matrix structure that needs to be estimated. Both sequential quadratic programming with channel-specific constraints (cSQP) and NMF were applied to estimate the spillover matrix using the cutoff separated spillover noise component. Finally, the CyTOF data can be compensated using the obtained spillover matrix. Figure 1A shows the workflow of our compensation method and Figure 1B describes the assumptions of our method.

3.1 | Simulation results

We applied two approaches, cSQP and NMF, to 20 different sets of simulated data. NMF did not reach final convergence, therefore, we took the results after 5000 iterations for evaluation purposes. In our simulation analysis, we compared both cSQP and NMF estimated spillover matrix to the simulated ground truths. We found that cSQP estimated results were better aligned to the ground truth than NMF. From the scatter plots it can be seen that most data points aligned on the diagonal in cSQP results which suggest that cSQP estimated similar spillover matrix as the simulated ground truths while the majority of NMF estimated spillover coefficients lied on the boundary of 0 and 0.1 (Figure 2A,B). The cSQP approach achieved $R^2 = 0.78$ compared to the NMF approach which achieved $R^2 = 0.02$. Since the cSQP approach can estimate spillover coefficients much more accurately than the NMF approach, we chose the cSQP approach for our downstream assessments.

The spillover effect is a property of the metal labels and the instrument and does not depend on biological markers or samples. We further demonstrated that the spillover effect is not affected by relative abundance of markers using simulation. We simulated data with different level of signals and obtained spillover effects that are independent of the level of signals (Figure S1). The detailed description of the simulation can be found in Supporting Information 1.

3.2 | Comparison with single-stained controls dependent compensation

We analyzed a PBMCs dataset stained using a 36-antibody panel with single-stained controls to compare the accuracy of the spillover matrices obtained using CytoSpill with those obtained using single-stained controls. In Figure 3A, we show the t-SNE plot of this data before compensation and it has 20 PhenoGraph clusters. The panel used for this data was developed for immune cell type identification, and some of the proteins on this

panel were conjugated with two different metal labels. In Figure 3B, we can observe the spillover intuitively by comparing the expression of the two metal labels on the same protein using expressing plots. The different expression between the two metals used for CD8, CD3 and HLA-DR suggested that there are spillovers in 174Yb-conjugated CD8, 173Yb-conjugated CD3 and 171Yb-conjugated HLA-DR. After compensation performed by CATALYST using single-stained controls, the same protein conjugated with different metals has almost identical expression profiles by comparing the expression plot. After compensation performed using CytoSpill generated spillover matrix, our method also removed the spillover effects and achieved similar results as using single-stained controls without requiring their use.

In Figure 3C, the heatmap of marker expressions in different clusters of uncompensated, single-stained controls compensated and CytoSpill compensated data can be seen. We annotated the clusters with cell types based on the marker expressions. We found that 173Yb-conjugated CD3 has expression in non-T cells clusters compared to 147Sm-conjugated CD3. The 174Yb-conjugated CD8 marker were observed to have expression in the non-CD8 T cells clusters compared to 139La-conjugated CD8. The 171Yb-conjugated HLA-DR had expressions in clusters besides the Macrophage/Monocytes and B cells clusters, while 175Lu-conjugated HLA-DR did not. The expression pattern in these clusters should be negative but were both removed using single-stained controls in CATALYST and ab initio in CytoSpill. This confirms that CytoSpill can effectively remove the true spillover in the CyTOF data without using single-stained controls.

3.3 | Applying CytoSpill on four immune datasets

Moreover, we also applied CytoSpill on four additional immune related datasets, including samples from human bone marrow, mouse bone marrow, human peripheral blood and human cord blood. We checked the marker expression before and after compensation. Also, t-SNE plots were generated and PhenoGraph was applied for clustering analysis [7, 17, 18]. We observed that some markers in these datasets have a high expression in certain clusters, while they also have some amount of intermediate expressions in other clusters. These intermediate expressions were lowered or removed after running CytoSpill.

For example, the CD8 markers in the leukemia patient peripheral blood data and the healthy human cord blood data were strongly affected by spillover (Figures 4 and 5). Figure 4A shows the t-SNE plot of uncompensated leukemia patient peripheral blood data, which is labeled by manually gated cell populations. We found that CD8 marker has an intermediate level in the B cells and CD4 T cells clusters, which may be caused by spillover. After performing CytoSpill compensation, these spillover signals were removed (Figure 4B,D). Figure 4C shows the histogram of the uncompensated and compensated arcsinh transformed CD8 expression with the dotted line representing the derived cutoff value for this channel. The noise component below the cutoff was substantially lowered after compensation with more cells having a level close to 0 and the positive population above the cutoff remained unchanged.

In the healthy human cord blood data, the CD8 marker has a high expression on the mucosal associated invariant T cell and the CD8 T cell clusters (Figure 5A,B). CD8 also expressed

on some of the NK cells. However, we found it also has intermediate levels on other clusters which should be negative for the CD8 marker. After performing compensation, the spillover caused signals were removed on these clusters (Figure 5B,D). Figure 5C showed the histogram compared the CD8 expression density before and after compensation. Noise component below the derived cutoff were lowered after compensation and the positive population above the cutoff remained.

3.4 | Discovering novel clusters

In terms of PhenoGraph clustering results, we found that the PhenoGraph clusters number increased from 22 to 23 after performing compensation on the healthy human bone marrow data, from 20 to 21 on the mouse bone marrow data and from 26 to 29 on the leukemia patient peripheral blood data. The increased number of clusters suggests that performing compensation may lead to discovery of novel cell clusters. In the healthy human bone marrow data, we found that two new clusters of T cells were revealed after compensation (Figure 6). Figure 6A–D shows the t-SNE plots for the clustering results obtained before and after compensation. Based on the t-SNE plots and PhenoGraph clustering results, two new clusters 2 and 4 from CD4 T cells and CD8 T cells respectively were formed after compensation (Figure 6C,D). By checking the expression signature of these two clusters we found that they are T cells with low CD44 expression which were not revealed as clusters before compensation (Figure 6E,F). CD44 is an activation marker which distinguishes naïve T cells from memory and effector T cells [19, 20]. These two novel clusters we found are thus likely naïve T cells clusters, which were masked by the spillover effects in the uncompensated data.

Identifying novel, meaningful clusters after performing compensation indicate that our compensation method can led to more precise cell-type clustering, in conjunction with the application of the t-SNE and the PhenoGraph algorithms.

4 | DISCUSSION

In immunology research, CyTOF was widely used to dissect the heterogeneity of immune cell populations. It is well accepted that the spillover effects in CyTOF data could led to inaccurate population dissection. In this article, we presented a new method that could alleviate spillover effects in CyTOF data without relying on additional control data such as single-stained controls. In our method, we used finite mixture model to derive the cutoffs that separates the noise and signal for each channel. We then use a constrained SQP approach to infer the spillover matrix that optimally quantifies the spillover effects. We performed simulation studies to demonstrate the ability of our method to deconvolve spillover effects on multiple published datasets and unpublished datasets. We observed markers affected by spillover effects and removed the noise after performing compensation. The compensation also led to increased number of PhenoGraph clusters in multiple datasets. In the healthy human bone marrow, our method discovered novel and meaningful cell subpopulations that would have been buried in the uncompensated data. To our knowledge this is the first method that can compensate spillover effects in CyTOF data without requiring single-stained controls.

We have considered two different approaches, cSQP and masked NMF, to integrate prior knowledge into our model and derive the spillover matrix. In our simulation study, we demonstrated that the cSQP approach had better performance than the NMF approach. NMF was not able to converge under our constraints. Since our spillover matrix is sparse under our assumption, some kind of sparsity constraint is required to be implemented in the optimization of NMF. There are several prior studies implement NMF based method on spectral unmixing, these studies have used sparsity constraints [21–23]. However, the spillover matrix structure was pre-defined based on prior knowledge in our study. To implement these prior knowledges in NMF, we used masked NMF [12–14]. Our constraint on the spillover matrix was too strict, leading to the undesirable performance of the masked NMF in our study.

Our method assumed that the spillover effects are caused by three main sources that are considered to be relatively mild (being a small fraction of the observed signals). Besides spillover effects, the noise in CyTOF data could also come from contaminations from other samples in the lab or external environment [8, 24]. However, these could be controlled based on rigorous experimental protocols [8]. Our method assumed the noise in the data come primarily from spillover. On datasets where these assumptions are violated (e.g., errors constitute a large fraction of the observed signal), the performance of our method may be impaired.

The current version of our method performed well on channels with high-intensity signals but worse on channels with low-intensity signals. Although this caveat may not affect the overall characterization results, which depend mainly on the high-intensity channels, users should be careful at interpreting coefficients estimated from low-intensity channels. In our future studies, we will investigate if the caveat can be addressed by performing further regularization or shrinkage that are inversely proportional to the intensity of the channels.

Our assumptions on spillover sources were based specifically on the CyTOF technology, which is different from the flow cytometry technology. Our method also assumed non-negativity of CyTOF data, while data from flow cytometry could have negative values due to background subtraction [25]. Thus, our method is effective for CyTOF data analysis and will not be applicable to flow cytometry data.

For the future work of our method, we would like to explore whether our proposed compensation method can lead to improvement in predicting clinical outcome and discovering novel disease mechanisms.

CytoSpill is implemented in R and the source code is released on GitHub: https://github.com/KChen-lab/CytoSpill. We expect that our method will significantly benefit the cancer and immunology research community in studying tumor microenvironment and developing novel immunotherapy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported in part by the University Cancer Foundation via the Institutional Research Grant program at the University of Texas MD Anderson Cancer Center, Human Breast Cell Atlas Seed Network Grants (CZF2019-002432 and CZF2019-02425) from the Chan Zuckerberg Initiative DAF, an advised fund of Sillicon Valley Community Foundation, grant RP180248 to Ken Chen from Cancer Prevention & Research Institute of Texas, a fellowship on the NLM Training Program T15LM007093 to Ramiz Iqbal, and P30 CA016672 (US National Institutes of Health/National Cancer Institute) to the University of Texas Anderson Cancer Center Core Support Grant.

Funding information

A training fellowship from the Gulf Coast Consortia, on the NLM Training Program in Biomedical Informatics & Data Science, Grant/Award Number: T15LM007093; Advised fund of Sillicon Valley Community Foundation; Cancer Prevention and Research Institute of Texas, Grant/Award Number: RP180248; Human Breast Cell Atlas Seed Network Grants from the Chan Zuckerberg Initiative DAF, Grant/Award Numbers: CZF2019-002432, CZF2019-02425; University Cancer Foundation via the Institutional Research Grant program at the University of Texas MD Anderson Cancer Center; US National Institutes of Health/National Cancer Institute, Grant/Award Number: P30 CA016672

REFERENCES

1. Alizadeh AA, Aranda V, Bardelli A, Blanpain C, Bock C, Borowski C, et al.Toward understanding and exploiting tumor heterogeneity. Nat Med. 2015;21(8):846–53. 10.1038/nm.3915. [PubMed: 26248267]
2. Wogsland CE, Greenplate AR, Kolstad A, Myklebust JH, Irish JM, Huse K. Mass cytometry of follicular lymphoma tumors reveals intrinsic heterogeneity in proteins including HLA-DR and a deficit in nonmalignant plasmablast and germinal center B-cell populations. Cytometry B Clin Cytom. 2017;92(1):79–87. 10.1002/cyto.b.21498. [PubMed: 27933753]
3. Kleinsteuber K, Corleis B, Rashidi N, Nchinda N, Lisanti A, Cho JL, et al.Standardization and quality control for high-dimensional mass cytometry studies of human samples. Cytometry A. 2016;89(10):903–13. 10.1002/cyto.a.22935. [PubMed: 27575385]
4. Roederer M. Spectral compensation for flow cytometry: visualization artifacts, limitations, and caveats. Cytometry. 2001;45(3):194–205. 10.1002/1097-0320(20011101)45:3<194::AID-CYTO1163>3.0.CO;2-C. [PubMed: 11746088]
5. Chevrier S, Crowell HL, Zanotelli VRT, Engler S, Robinson MD, Bodenmiller B. Compensation of signal Spillover in suspension and imaging mass Cytometry. Cell Syst. 2018;6(5):612–620.e5. 10.1016/j.cels.2018.02.010. [PubMed: 29605184]
6. Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype space with single-cell data. Nat Methods. 2016;13:493–6. [PubMed: 27183440]
7. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, et al.Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015;162(1):184–97. 10.1016/j.cell.2015.05.047. [PubMed: 26095251]
8. Takahashi C, Au-Yeung A, Fuh F, Ramirez-Montagut T, Bolen C, Mathews W, et al.Mass cytometry panel optimization through the designed distribution of signal interference. Cytometry A. 2017;91(1):39–47. 10.1002/cyto.a.22977. [PubMed: 27632576]
9. Roederer M. Compensation in flow cytometry. Curr Protoc Cytom. 2002; 22(1):1.14.1–1.14.20. 10.1002/0471142956.cy0114s22.
10. Grün B, Leisch F. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. J Stat Softw. 2008;28(4). 10.18637/jss.v028.i04.
11. Kraft DA. Software package for sequentml quadratic programming. technical report DFVLR-FB 88–28, Oberpfaffenhofen: Institut fuer Dynamik der Flugsysteme, 1988.
12. Casalino G, Del Buono N, Mencar C. Nonnegative matrix factorizations for intelligent data analysis. In: Naik GR, editor. Non-negative matrix factorization techniques: advances in theory and applications. Berlin, Heidelberg: Springer; 2016. p. 49–74. 10.1007/978-3-662-48331-2_2.
13. Lin X, Boutros PC. Optimization and expansion of non-negative matrix factorization. BMC Bioinformatics2020;21(1):7. 10.1186/s12859-019-3312-5. [PubMed: 31906867]

14. Sobieraj I, Kong Q, Plumbley MD. 2017. Masked non-negative matrix factorization for bird detection using weakly labeled data. Paper presented at: 25th European Signal Processing Conference (EUSIPCO); 2017, p. 1769–1773. 10.23919/EUSIPCO.2017.8081513
15. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al.2015. TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/
16. Chen D, Plemmons RJ. Nonnegativity constraints in numerical analysis. In: Bultheel A, Cools R, editors. Symposium on the birth of numerical analysis. Singapore: World Scientific Press; 2009.
17. Krijthe J. 2020. jkrijthe/Rtsne [C++]. https://github.com/jkrijthe/Rtsne
18. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(Nov):2579–605.
19. Baaten BJ, Li C-R, Bradley LM. Multifaceted regulation of T cells by CD44. Commun Integr Biol. 2010;3(6):508–12. 10.4161/cib.3.6.13495. [PubMed: 21331226]
20. Schumann J, Stanko K, Schliesser U, Appelt C, Sawitzki B. Differences in CD44 surface expression levels and function discriminates IL-17 and IFN-γ producing helper T cells. PLoS One. 2015;10(7):e0132479. 10.1371/journal.pone.0132479. [PubMed: 26172046]
21. Jiménez-Sánchez D, Ariz M, Morgado JM, Cortés-Domínguez I, Ortiz-de-Solórzano C. NMF-RI: blind spectral unmixing of highly mixed multispectral flow and image cytometry data. Bioinformatics. 2020;36(5):1590–8. 10.1093/bioinformatics/btz751. [PubMed: 31593222]
22. Lu X, Wu H, Yuan Y. Double constrained NMF for hyperspectral unmixing. IEEE Trans Geosci Remote Sens. 2014;52(5):2746–58. 10.1109/TGRS.2013.2265322.
23. Rajabi R, Ghassemian H. Spectral unmixing of hyperspectral imagery using multilayer NMF. IEEE Geosci Remote Sens Lett. 2015;12(1):38–42. 10.1109/LGRS.2014.2325874.
24. Leipold MD, Newell EW, Maecker HT. Multiparameter phenotyping of human PBMCs using mass cytometry. Methods Mol Biol (Clifton, NJ). 2015;1343:81–95. 10.1007/978-1-4939-2963-4_7.
25. Tung JW, Heydari K, Tirouvanzia R, Sahaf B, Parks DR, Herzenberg LA, et al.Modern flow cytometry: a practical approach. Clin Lab Med. 2007;27(3):453–v. 10.1016/j.cll.2007.05.001. [PubMed: 17658402]

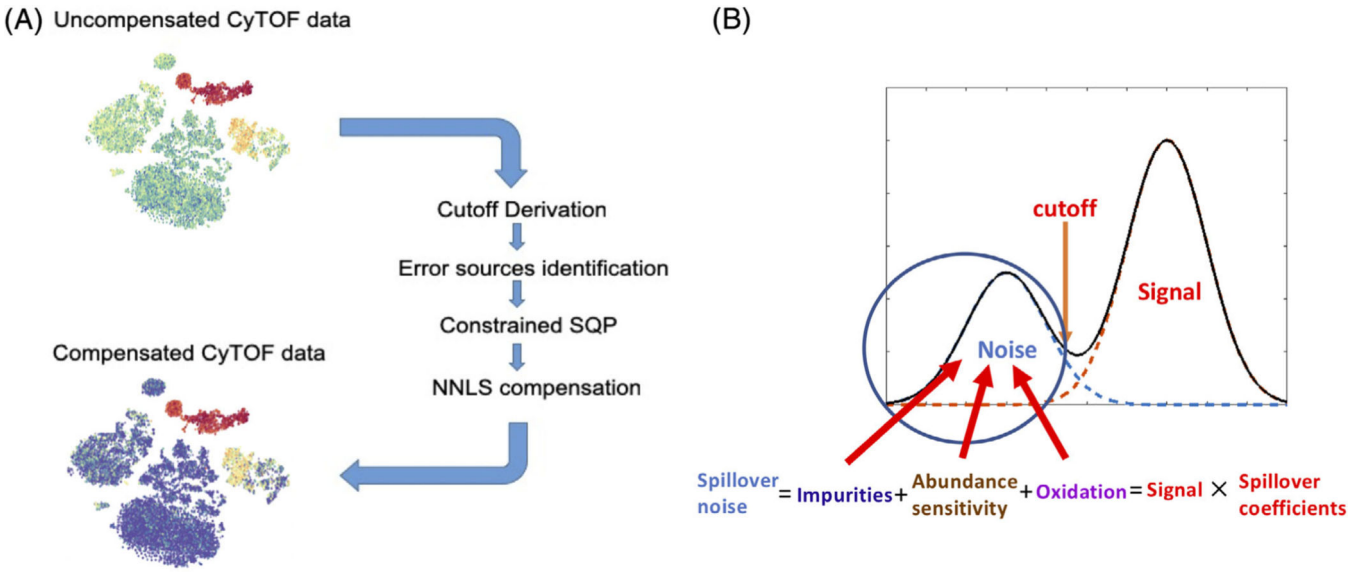

FIGURE 1.

(A) Workflows of our compensation method. Based on our assumptions, we derived cutoffs based on uncompensated mass cytometry (CyTOF) data to identify the spillovers and estimate the spillover matrix using constrained sequential quadratic programming with prior knowledge. Non-negative least squares were used for compensation. (B) Assumptions of our methods. We assumed that a spillover affected channel will have a lower modal which contributed by spillover effects. We assumed the spillover was from three sources: Isotope impurities, neighboring channel abundance sensitivity and oxidization

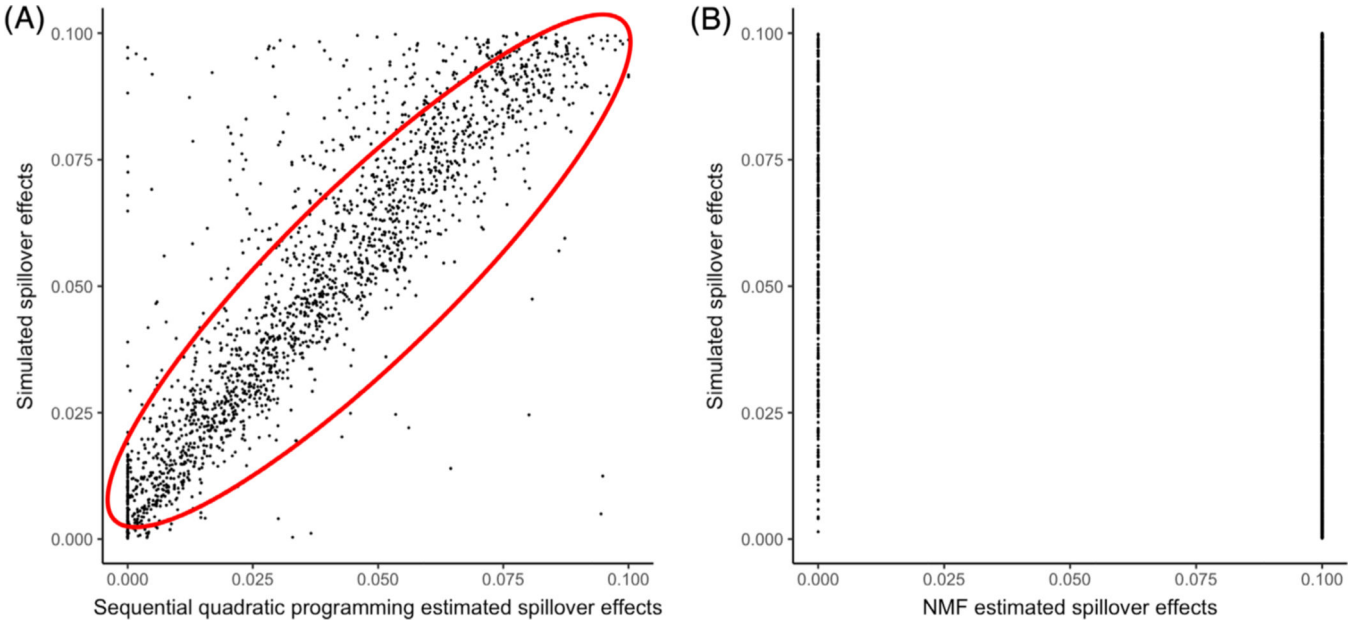

FIGURE 2.

Comparison of simulated spillover effects and estimated spillover effects. X-axis represented the estimated spillover effects using our methods. Y-axis represented the simulated true spillover effects. Each dot in the figure represented an entry in a spillover matrix of our simulation study. (A) Scatter plot shows the result of sequential quadratic programming estimation and the $R^2 = 0.78$. (B) Scatter plot shows the result of non-negative matrix factorization and the $R^2 = 0.02$

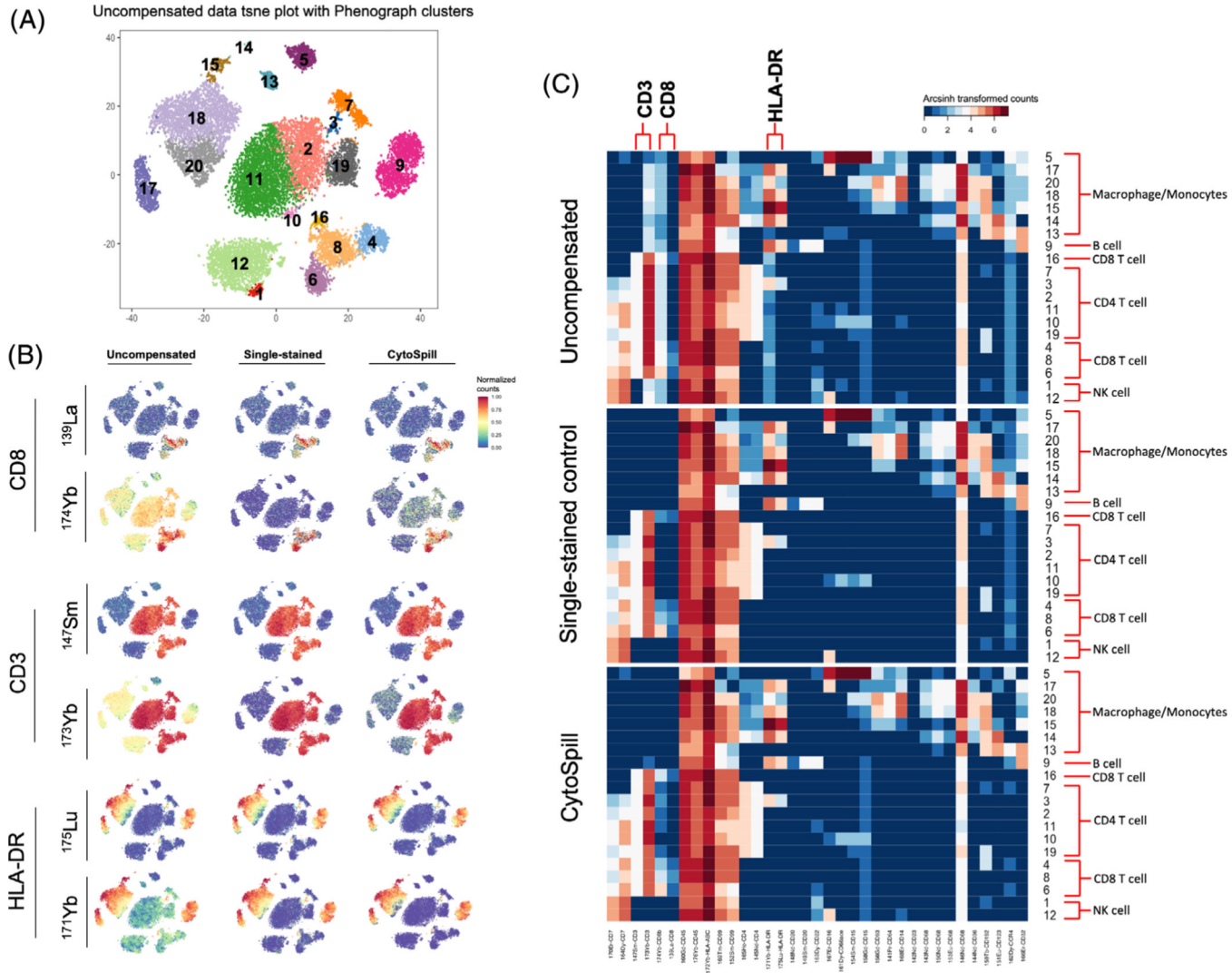

FIGURE 3.

Comparison of CytoSpill generated spillover matrix compensated data with single-stained controls generated spillover matrix compensated data. (A) The t-SNE plot of uncompensated data labeled with PhenoGraph clusters. (B) Marker expressions based on the t-SNE projection by uncompensated data. Expression values were normalized between 0 and 1. Spillovers were observed in 174Yb-stained CD8, 173Yb-stained CD3 and 171Yb-stained HLA-DR. CytoSpill achieved comparable results with single-stained controls on compensating these markers. (C) Heatmaps showed compensation results on PhenoGraph clusters based on uncompensated data. Expression values were arcsinh transformed. Clusters were annotated with cell types

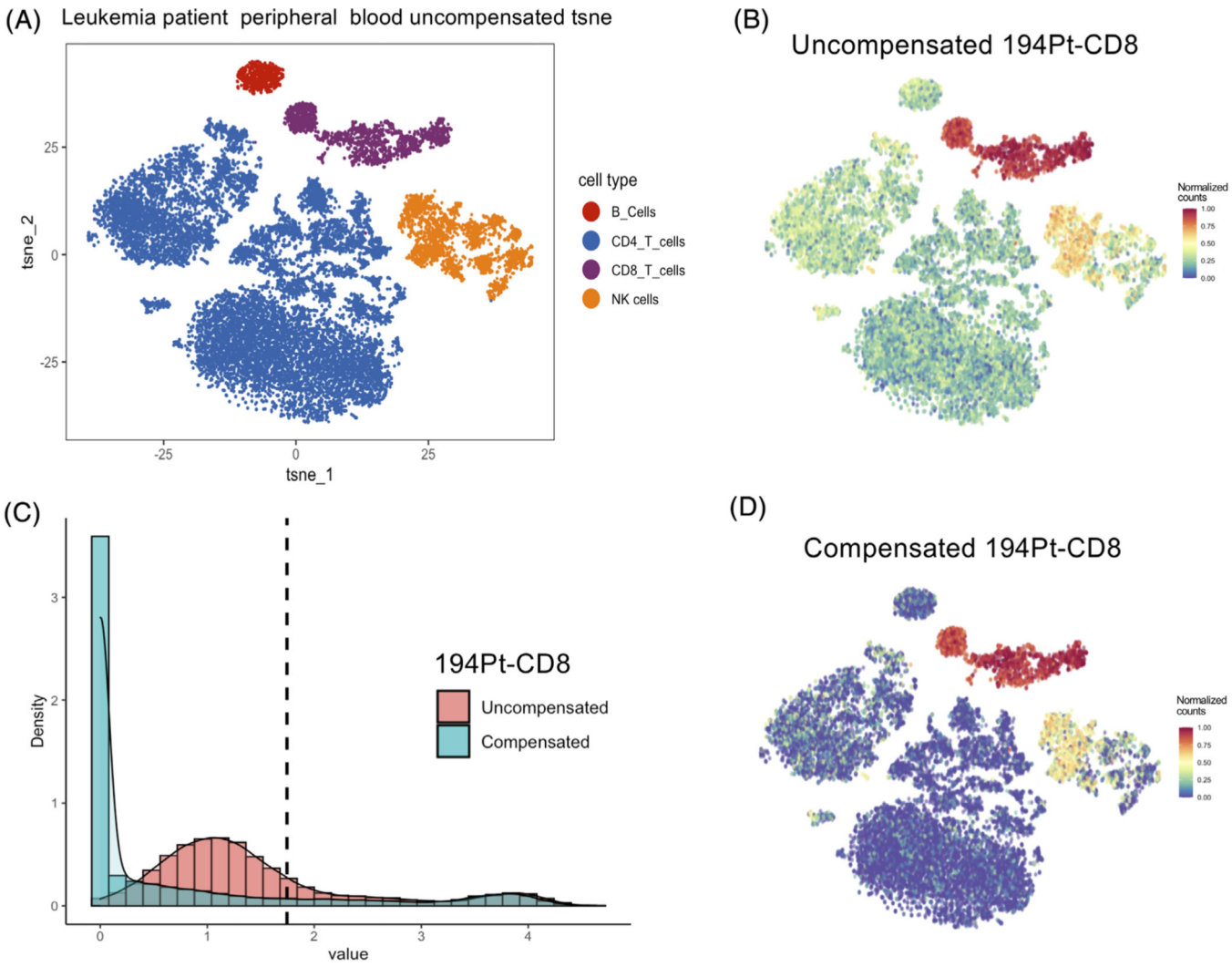

FIGURE 4.

CD8 marker spillover in leukemia patient blood data were compensated using CytoSpill. (A) The t-SNE plot generated with uncompensated data and labeled by manually gated populations. (B) The normalized expression plot of the 194Pt-conjugated CD8 marker before compensation. Spillover was observed outside NK cells and CD8 T cells. (C) Histograms showed the CD8 marker expression density before and after compensation. The dotted line represents the derived cutoff value. The noise component below the cutoff was lowered after compensation. (D) The normalized expression plot of the 194Pt-conjugated CD8 marker after compensation. We observed that spillover was lowered after compensation

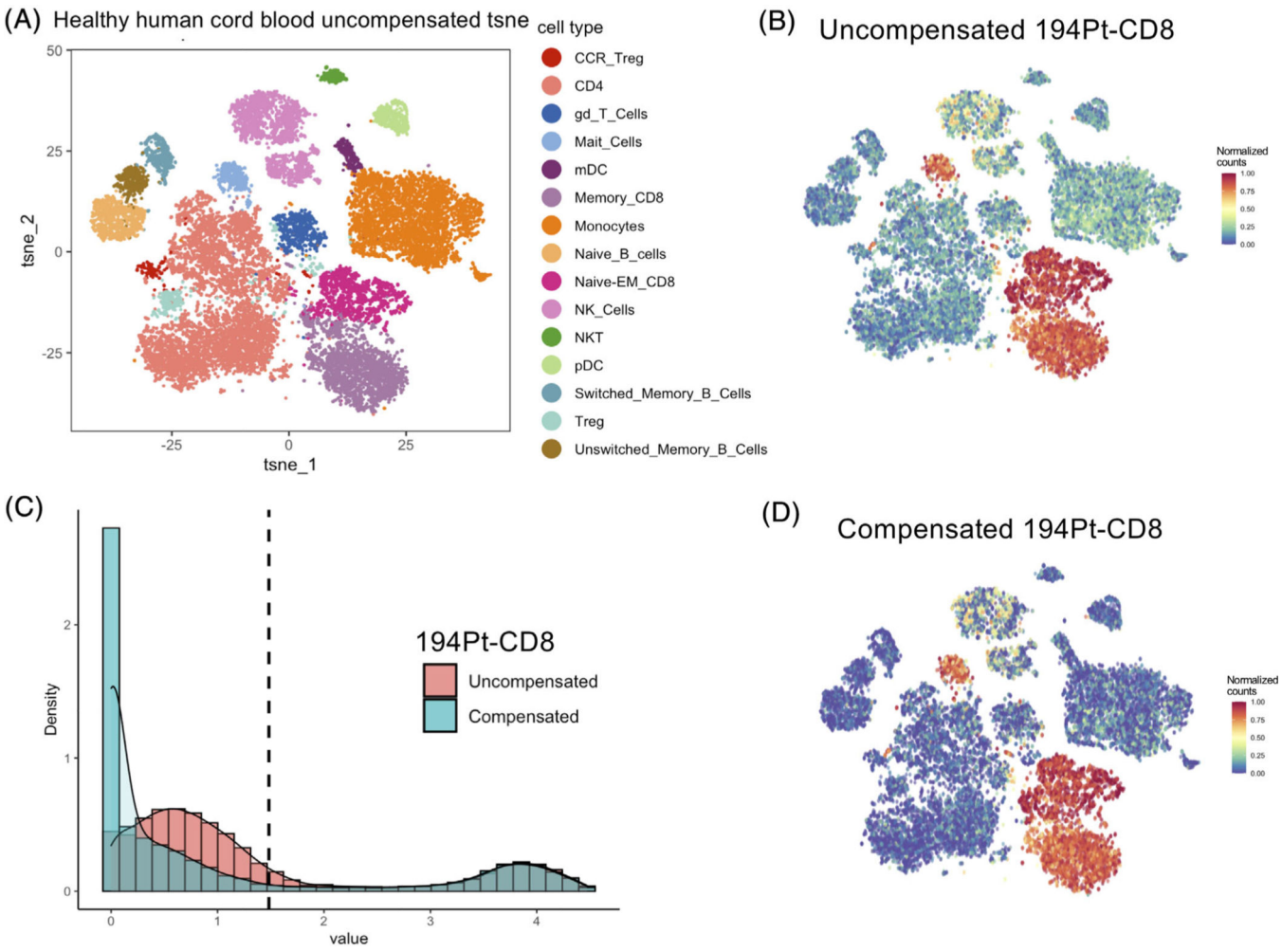

FIGURE 5.

CD8 marker spillover in healthy human cord blood data were compensated using CytoSpill. (A) The t-SNE plot generated with uncompensated data and labeled by manually gated populations. (B) The normalized expression plot of the 194Pt-conjugated CD8 marker before compensation. Spillover was observed outside CD8 T cells, NK cells and Mait cells. (C) Histograms showed the CD8 marker expression density before and after compensation. The dotted line represents the derived cutoff value. The noise component below the cutoff was lowered after compensation. (D) The normalized expression plot of the 194Pt-conjugated CD8 marker after compensation. We observed that the spillover was lowered after compensation

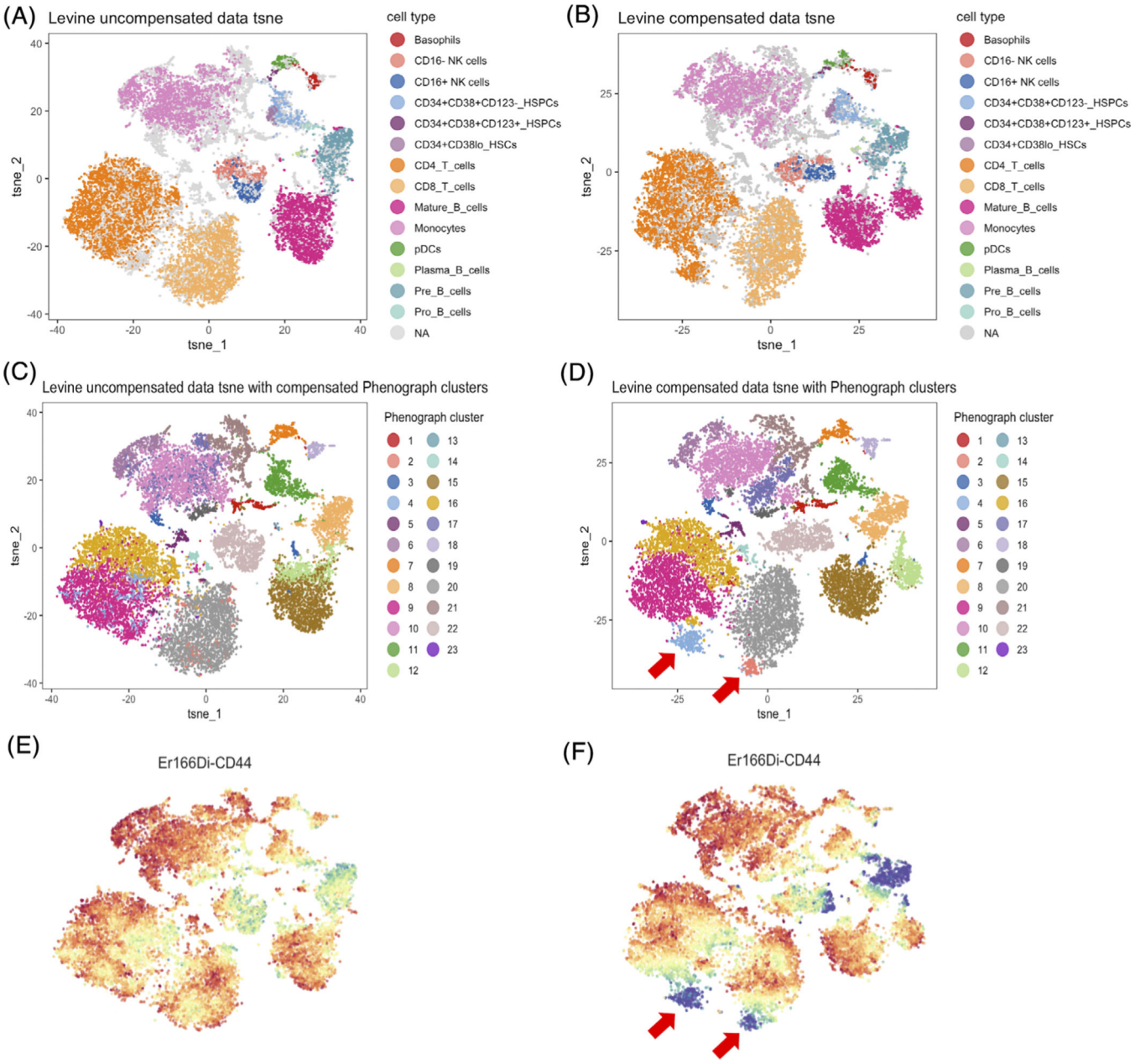

FIGURE 6.

Novel clusters discovered after compensation in healthy human bone marrow data. (A, B) Showed the t-SNE plots generated based on uncompensated and compensated data labeled by cell populations. (C, D) Showed the t-SNE plots labeled by compensated data generated PhenoGraph clusters. Two new clusters 2 and 4 were found after compensation. These two clusters of cells were originally scattered in CD4 T cells and CD8 T cells before compensation. (E, F) Showed the CD44 expression level before and after compensation. The newly found clusters have lower CD44 expression

TABLE 1

Descriptions of the datasets used in this study

Datasets	Source	Number of antibody channels	Number of cells	Tissue
Peripheral blood mononuclear cells	[5]	36	87,817	Human peripheral blood mononuclear cells
C57BL/6J mouse bone marrow	[6]	38	86,864	Mouse bone marrow
Healthy human bone marrow	[7]	32	265,627	Human bone marrow
Chronic lymphocytic leukemia patient blood	New	46	126,855	Human peripheral blood
Healthy human cord blood	New	44	173,891	Human cord blood

TABLE 2

Commonly used notations in this article

Notations	Definition
$\boldsymbol{D}$	Matrix of observed signals
$\boldsymbol{T}$	Matrix of true signals
$\boldsymbol{S}$	Spillover matrix
$\boldsymbol{Y}$	Noise component in the matrix of observed signals
$\boldsymbol{Z}$	Channel interaction matrix
$\boldsymbol{I}$	Identity matrix
N	Number of cells
M	Number of channels
$\|\cdot\|^2_F$	Frobenius norm of a matrix