Cytometry Part A

CYTOMETRY Journal of Quantitative Cell Science PART A

ORIGINAL ARTICLE

Visual Quality Control With *CytoMDS*, a Bioconductor Package for Low Dimensional Representation of Cytometry Sample Distances

Philippe Hauchamps[1] | Simon Delandre[2] | Stéphane T. Temmerman[2] | Dan Lin[2] | Laurent Gatto[1]

[1]Computational Biology and Bioinformatics, de Duve Institute UCLouvain, Woluwe-Saint-Lambert, Belgium | [2]GSK, Rixensart, Belgium

Correspondence: Laurent Gatto (laurent.gatto@uclouvain.be)

Received: 8 July 2024 | **Revised:** 9 December 2024 | **Accepted:** 30 January 2025

Funding: This work was supported by GlaxoSmithKline Biologicals S.A., under a cooperative research and development agreement between GlaxoSmithKline Biologicals S.A. and de Duve Institute (UCLouvain).

ABSTRACT

Quality Control (QC) of samples is an essential preliminary step in cytometry data analysis. Notably, the identification of potential batch effects and outlying samples is paramount to avoid mistaking these effects for true biological signals in downstream analyses. However, this task can prove to be delicate and tedious, especially for datasets with dozens of samples. Here, we present *CytoMDS*, a Bioconductor package implementing a dedicated method for low-dimensional representation of cytometry samples composed of marker expressions for up to millions of single cells. This method allows a global representation of all samples of a study, with one single point per sample, in such a way that projected distances can be visually interpreted. *CytoMDS* uses *Earth Mover's Distance* for assessing dissimilarities between multi-dimensional distributions of marker expression and *Multi-Dimensional Scaling* for low-dimensional projection of distances. Some additional visualization tools, both for projection quality diagnosis and for user interpretation of the projection coordinates, are also provided in the package. We demonstrate the strengths and advantages of *CytoMDS* for QC of cytometry data on three real biological datasets, revealing the presence of low-quality samples, batch effects, and biological signal between sample groups.

1 | Introduction

Recent technology advances in flow and mass cytometry allow the acquisition of high-dimensional sample data, with up to 40 and 50 parameters measured simultaneously for millions of cells, respectively [1, 2]. For such high-dimensional data samples, extracting relevant biological information becomes challenging [3]. Moreover, in large clinical studies, possibly with some longitudinal and multicentric components, this difficult task is itself multiplied by the high number of samples [4, 5]. In the last decade, this led to the development of (semi) automated data analysis pipelines [6] to complement the traditional manual gating approaches [7]. Among the different steps involved in (automated) data analysis pipelines, pre-processing and Quality Control (QC) of data samples is paramount, as it guarantees that the downstream analyses are not compromised by confounding noise. For cytometry data, pre-processing and QC involve different treatments and checks [8]: compensation, stability of signal in time, removal of undesirable events like debris, doublets, or dead cells, batch effect detection and removal, and so forth.

Whatever the level of automation of these treatments, some manual checks of the cleaned samples, possibly at different stages of the pre-processing, are still necessary. This is typically done by human operators using interactive visual software, like for example FlowJo or Cytobank—among the available commercial platforms, or free open source solutions [9, 10]. Nevertheless, these tools are mostly designed to check samples one by one or possibly to visually compare two or a handful of samples at once. However, two categories of QC, namely outlying sample identification [11] and batch

effect detection [12], would benefit from a global representation of the dataset. Well-known classical data exploration techniques like *Principal Component Analysis (PCA)* [13] have the potential to provide such a global view of the data and are routinely used in other omics contexts like for example bulk transcriptomics [11] and bulk proteomics [14] data analyses. In bulk experiments, each sample results in a single vector of measurements because all its cells are merged. All samples can be combined into a single matrix which corresponds to the use of a standard PCA or a similar algorithm. In single cell experiments, there is a measurement vector for each cell and each sample results in a measurement matrix. Applying these bulk techniques to cytometry data presents a major challenge in summarizing each sample by a vector for comparing samples (not their cells) with each other.

To address this challenge, we developed *CytoMDS* [15], an R package for low-dimensional projection of cytometry samples, which allows a global representation of all samples of a study, with one single point per sample, in such a way that projected distances can be visually interpreted. In this article, we present the *CytoMDS* method and demonstrate its use on three real cytometry datasets, illustrating the detection of various quality issues. Finally, we compare our approach with similar existing methods for global problem detection in cytometry samples and discuss its advantages and potential limitations. *CytoMDS* is available in Bioconductor [16], as of version 3.19.

2 | Materials and Methods

2.1 | Proposed Method

Figure 1 sketches an overview of the *CytoMDS* workflow. As input, the method requires a set of minimally pre-processed—that is, scale transformed and, if applicable, compensated or unmixed—sample files in *flow cytometry standard (fcs)* format (Figure 1A). The *CytoMDS* method itself (Figure 1B) can be decomposed into two main steps:

- computing the distances between each sample pair (1);
- projecting the resulting pairwise distance matrix onto a low-dimensional space (2) of dimension p, p being typically 2, 3, or a small number sufficient to reach a good projection quality (see below).

As a result, each sample is a single point in this p-dimensional space, and all samples can be represented in plots combining those p dimensions. Optionally, *bi-plots* can be performed by overlaying on these plots the correlations (4) of chosen sample statistics (3) with the projection coordinates in order to interpret the directions within the plots.

2.1.1 | Distance Calculation

In order to characterize the dissimilarity between samples, the *Earth Mover's Distance (EMD)*, a.k.a. the *Wasserstein distance of order one*, is chosen as our approach. Considering two samples a and b, which are represented by their respective M-dimensional distributions of expression intensities for the M markers, the *EMD* can be seen as the minimal cost (or effort) needed for transporting the distribution mass of one sample to the other, where the cost of transporting one unit of mass from location x to location y is given by the Euclidean distance between x and y [17].

The *EMD* has been shown to be suitable for quantifying dissimilarities between cytometry samples and to provide better insight than other types of metrics for this purpose [18]. It has been used in different applications of cytometry data analysis, especially in the evaluation of batch correction methods [19, 20] or for obtaining a sample representation in relation to different phenotypes [21].

Here, for computational efficiency given the potentially high number of markers, we approximate the multi-dimensional *EMD* by the sum of the M 1D *EMD* for each marker m: 1, M. Assuming that each unidimensional marker expression distribution has been discretized in 1D histograms consisting of a number of bins with specific location and weight, the 1D *EMD* is then formally defined as follows:

- Let a_i and w_i^a (i: {1, … r}) be respectively the location and weight assigned to bin i of the 1D histogram of marker m in sample a.
- Let b_j and w_j^b (j: {1, … s}) be respectively the location and weight assigned to bin j of the 1D histogram of marker m in sample b.

The empirical distribution functions $F(t)$ and $G(t)$ of the marker m intensity, respectively for sample a and b, are therefore: $F(t) = \sum_{i=1}^{r} w_i^a 1\{a_i \le t\}$ and $G(t) = \sum_{j=1}^{s} w_j^b 1\{b_j \le t\}$.

The 1D *EMD* between marker m intensities of samples a and b is then:

$$W_1(F, G) = \int_{-\infty}^{\infty} | F(z) - G(z) | \, dz \tag{1}$$

To compute the *EMD* for each marker (Equation (1)), we take advantage of the R package *transport* [22]. This computation is then repeated for each pair of samples, leading to a symmetric pairwise distance matrix having N rows and N columns—with N being the number of samples. Its diagonal is filled with zeros.

2.1.2 | Projection Algorithm

The second step consists of projecting the distance matrix of dimensions NxN onto a low-dimensional embedding with p dimensions. As our goal is to visualize sample distances to identify outliers or batch effects, we chose a method from the *Multi-Dimensional Scaling (MDS)* family, since they are good at preserving distances and the global data structure [23], in contrast to popular local neighborhood-preserving methods such as *t-SNE* [24] and *UMAP* [25].

Furthermore, because the *EMD* is a metric but is not euclidean, we need to resort, for the projection, to *non-classical metric MDS*. It is a non-linear method minimizing a specific criterion called *stress* [23] that represents a quadratic error of projection and is defined as follows:

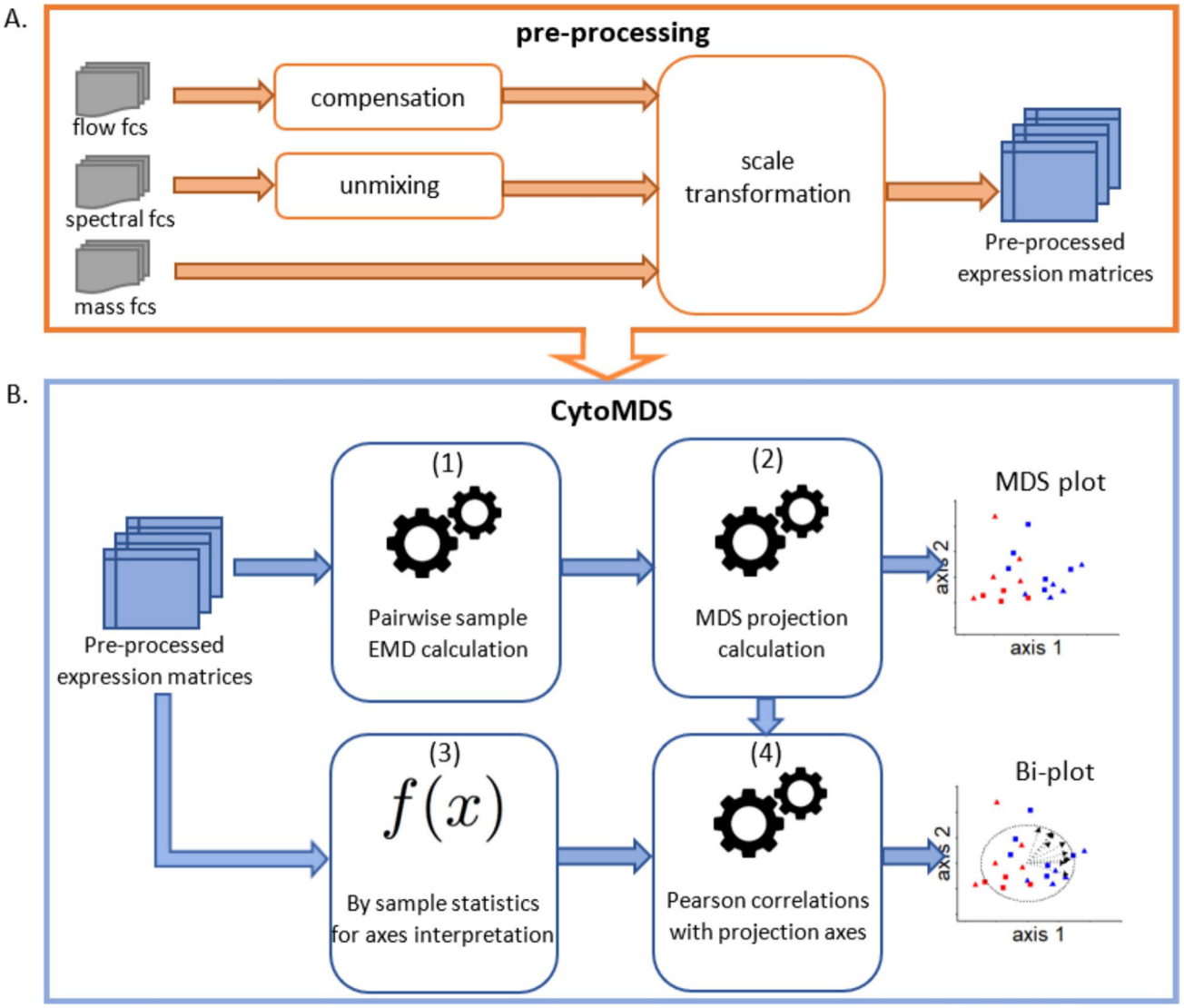

FIGURE 1 | Overview of the *CytoMDS* workflow. Pre-processing of the raw *fcs* files (A) should be done according to the workflow in the orange box, which includes scale transformation and might include compensation or unmixing depending on the type of data. The *CytoMDS* method itself (B) takes as input one pre-processed expression matrix per sample. First, *Earth Mover's Distances (EMD)* are calculated for each pair of samples (1). Second, the resulting distance matrix is projected using *Multi Dimensional Scaling (MDS)* (2), producing an *MDS* plot. Optionally, in order to interpret *MDS* coordinates and plot directions, *CytoMDS* can compute a set of chosen sample statistics (3). *CytoMDS* then calculates Pearson correlations of these statistics with the *MDS* coordinates (4), producing one or several bi-plot(s). [Color figure can be viewed at wileyonlinelibrary.com]

$$\sigma(\boldsymbol{X}) = \sum_{i<j} \left(\delta_{ij} - d_{ij}(\boldsymbol{X})\right)^2 \tag{2}$$

where $\boldsymbol{X}$ is the low-dimensional embedding, δ_{ij} is the original *EMD* between sample i and sample j, and d_{ij} is the euclidean distance between the corresponding projections in $\boldsymbol{X}$.

In order to minimize the stress criterion (Equation (2)), *CytoMDS* uses the *SMACOF* (Scaling by MAjorizing a COmplicated Function) algorithm, implemented in the *smacof* R package [26]. Finally, the projection axes of the p-dimensional embedding are linearly transformed using *PCA*. This last step results in new projection coordinates that are ordered in decreasing percentages of variance, as in a classical *PCA* visualization. However, here these percentages are with respect to the total variance associated with the embedding, which differs from the total variability of the data.

2.1.3 | Quality of Projection

CytoMDS provides different diagnostic tools to assess the quality of the low dimensional projection. The first one is a visualization plot, called *Shepard's diagram*, which allows to assess how close the projected euclidean distances are to their *EMD*, for each sample pair. Examples of such diagrams can be found in Figures S1 and S3. The second diagnostic tool is a global indicator of the projection quality, the *pseudo* R^2, which assesses to which extend the pairwise sample *EMD* can be explained by the low dimensional euclidean distances on the projection. It is defined as follows:

$$pseudo R^2 = 1 - \frac{\sum_{i<j}\left(\delta_{ij} - d_{ij}(\boldsymbol{X})\right)^2}{\sum_{i<j}\left(\delta_{ij} - \bar{\delta}\right)^2}, \quad (3)$$

where $\bar{\delta}$ is the overall mean of pairwise *EMD*.

This *pseudo* R^2 plays a central role in *CytoMDS* as, by default, p, the number of dimensions of the low dimensional embedding, is tuned automatically to reach a specific *pseudo* R^2 target for the projection, 0.95 by default. Finally, additional indicators from the *SMACOF* algorithm can also be recovered, such as the minimized value of the stress criterion (Equation (2)) and the contribution of each sample to the stress, that is the *stress per point* [26].

2.1.4 | Coordinates and Directions Interpretation Using Bi-Plots

Bi-plots are plots that are overlaid on the *MDS* projection, allowing to interpret directions of interest by relating them to well-chosen sample characteristics [27]. Creating a bi-plot involves choosing *external scales*, that is, sample-specific characteristics that can be fitted to the *MDS* space. Possible choices are summary statistics of marker intensities, for example univariate mean, standard deviation, median, or other quantiles, and so forth, or any other sample-representative characteristics (multivariate statistics, number of events in sample, etc.). *CytoMDS* calculates the Pearson correlations of the *external scales* with respect to the visualized *MDS* coordinates and overlays arrows on the plot, one for each *external scale*, having coordinates equal to their respective correlations. A unit 1 correlation circle is also plotted as a landmark for visual convenience (e.g., in Figure 2C,D).

2.2 | Example Datasets

In order to demonstrate *CytoMDS* in subsequent sections, we make use of three datasets.

2.2.1 | HBV Chronic Mouse Dataset

The *HBV Chronic Mouse* dataset is a flow cytometry dataset, collected in a preclinical study aimed at assessing the effect of different therapeutic vaccine regimens on the immune response of Hepatitis B Virus transduced mice. This dataset has been described in [10]. In short, it contains 55 flow cytometry mouse liver lymphocyte samples, analyzed with a 12-channel flow cytometry panel. The 55 mice were split into 5 experimental groups corresponding to different vaccine regimens, and the flow cytometry data were acquired in two flow cytometry batches performed on two different days (D91 and D93). The *HBV Chronic Mouse* dataset is available on Zenodo (DOI:10.5281/zenodo.8425840).

2.2.2 | ImmunoSenescence Human PBMC Dataset

The *ImmunoSenescence Human PBMC* dataset is also a flow cytometry dataset. This dataset has been generated in the context of the development of a 26-marker panel that aimed at phenotyping all the main immune cell populations present in PBMC (Peripheral Blood Mononuclear Cell) samples. This readout development was part of a study that aimed at assessing the effect of age on the human immune system. The dataset contains 20 *fcs* files, corresponding to the flow cytometry data acquisition of 15 PBMC samples. These were collected and cryopreserved from healthy individuals, split into two age groups—younger and older adults—and were part of two different GSK biobanks. There is therefore a potential confounding effect between the biological signal of interest—linked to aging—and the two different sample origins. The flow cytometry data acquisition itself was performed in two different acquisition rounds (a *former* and a *later* data acquisition), and 5 of the biological samples were acquired in both rounds (see Table S1). Sampling and research were approved by the Belgian Ethics Committee, and written informed consent was obtained from all participants prior to sample collection, in accordance with the Helsinki Declaration.

2.2.3 | Krieg_Anti_PD1 Dataset

The *Krieg_Anti_PD1* dataset is a mass cytometry dataset [28], consisting of 20 baseline samples (prior to treatment) of peripheral blood from melanoma skin cancer patients subsequently treated with anti-PD-1 immunotherapy. The samples were split across 2 conditions (non-responders and responders) and into 2 acquisition batches. The goal of the study was to identify biomarkers of responsiveness to immunotherapy at baseline. This dataset is available in FlowRepository (ID: FR-FCM-ZYL8) and is directly accessible in R from the Bioconductor data package *HDCytoData* [29].

3 | Results

3.1 | Identifying Low Quality Samples

In this section, we demonstrate the use of *CytoMDS* on the *HBV Chronic Mouse* dataset, with QC as a specific goal. The raw data were pre-processed by applying compensation and scale transformation using, respectively, bi-exponential transformations [30] for fluorescent channels and linear transformations for *FSC* and *SSC* channels. Coefficients of the linear transformations were calibrated to align the 5 and 95 quantiles of *FSC/SSC* channels to those of a reference fluorescent channel. *CytoMDS* was used to compute *EMD* between each pair of samples. Here, since the focus is on tracking sample quality issues, we limited ourselves to four specific channels as input for the *EMD* calculation, that is, *FSC-A*, *FSC-H*, *SSC-A*, and the *Live/Dead* channels, which are the channels used for pre-gating [31]. Finally, the low dimensional projection was computed for the 55 samples of the dataset.

The *Shepard's diagram*, in Figure S1, illustrates the projection quality. It shows that most pairwise *EMD* are close to the 45° identity line. Also, the *pseudo* R^2 is 0.9679, hence above the 0.95 threshold (chosen by default), with only two projected dimensions.

The projection, in Figure 2A, highlights two major clusters of samples. On the positive side of the x coordinate, we can find most of the samples acquired on day D93 (blue squares). The other cluster of samples, on the negative side of the x coordinate, includes mostly the samples from day D91 (red triangles), but also a minority of day D93 samples. Therefore,

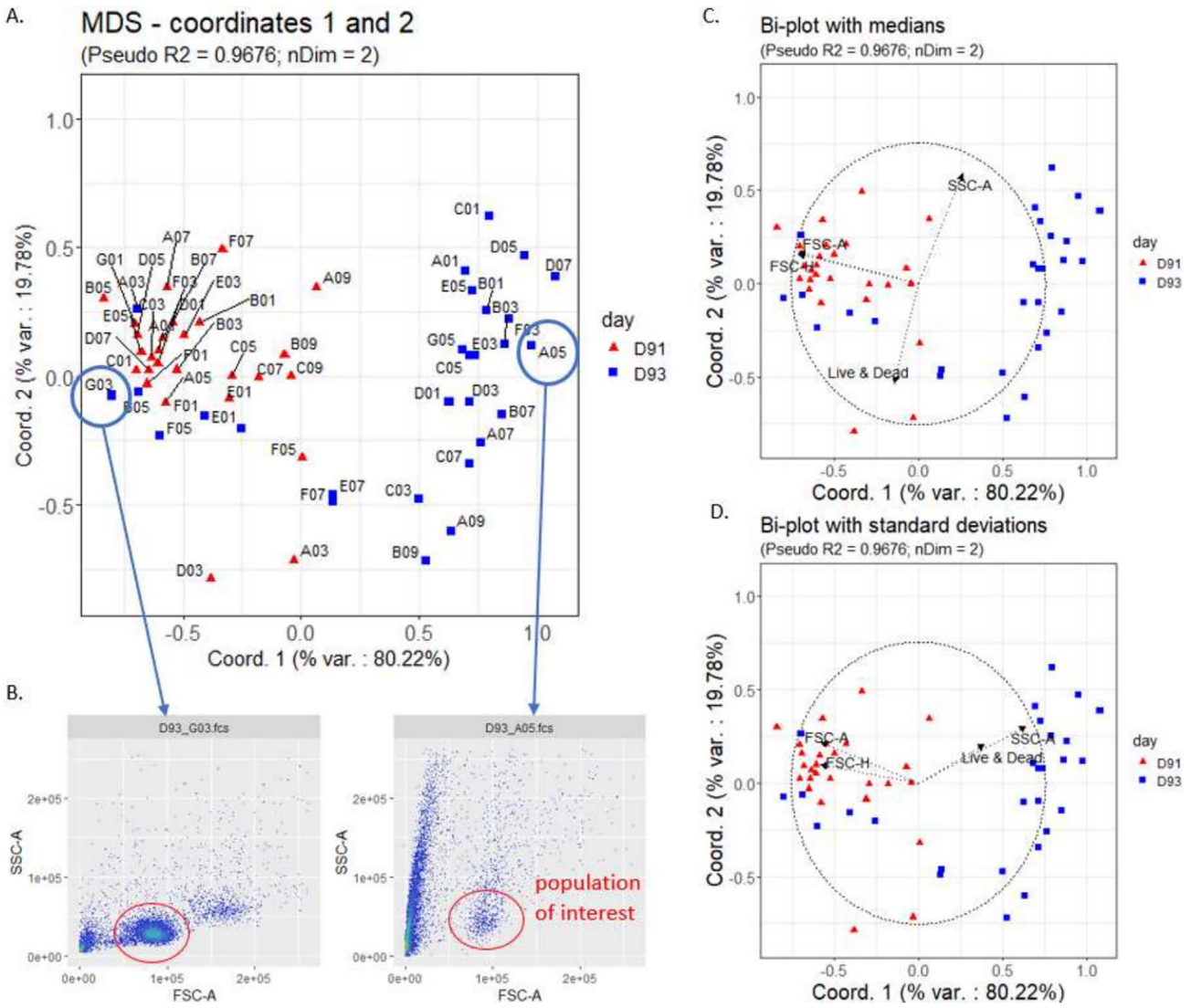

FIGURE 2 | The low-dimensional projection of the *HBV Chronic Mouse* dataset (A) reveals a clear split between two sample groups, not fully attributable to a batch effect linked to the day of acquisition—since a minority of day D93 samples belong to the left sample group. Bi-plots overlaying respectively channels medians (C) and standard deviations (D) show that the x axis, which is the direction discriminating the two sample groups, is strongly negatively correlated both with *FSC* channel median and *FSC* standard deviation, and also positively correlated with *SSC* standard deviation. A comparison of the *FSC* versus *SSC* 2D representation of selected samples having opposite x coordinates (B) shows that the right sample group contains low-quality samples, with very few events belonging to the cell population of interest. [Color figure can be viewed at wileyonlinelibrary.com]

these clusters cannot be attributed solely to a batch effect of acquisition time point. Two bi-plots using medians and standard deviations (Figure 2, respectively [C] and [D]) show that the x coordinate is strongly correlated, in a negative manner, with both the *FSC-A* channel median and *FSC-A* standard deviation, and positively correlated with *SSC-A* standard deviation. *FSC-A* and *SSC-A* are therefore taken as plot axes in Figure 2B, showing 2D plots for two representative samples, selected for their extreme opposite position on the first axis of the projection. Comparing these two plots reveals that sample *D93_A05* is a low-quality sample where the cell population of interest—liver lymphocytes—is extremely small, and where most events seem to correspond to debris or dead/dying cells, whereas sample *D93_G03* is a good-quality sample. Actually, as shown in Figure S2, the same pattern can be seen in all samples belonging to the low-quality cluster.

As a conclusion, the *CytoMDS* projection of the *HVB Chronic Mouse* dataset, complemented by bi-plots, has allowed identifying a sample cluster corresponding to low-quality samples, in terms of relative percentage of events belonging to the cell population of interest.

3.2 | Highlighting Batch Effects

In this section, we apply *CytoMDS* on the *ImmunoSenescence Human PBMC* dataset. As in the previous section, we preprocessed the data by applying compensation—based on a manually adjusted compensation matrix—and scale transformed the channel intensities—using bi-exponential transformations or linear transformations as in previous section. The pairwise sample *EMD* were calculated using the *FSC-A*, *FSC-H*, *SSC-A*

and *Live/Dead* channels. In contrast to the *HBV Chronic Mouse* dataset, *CytoMDS* required 3 dimensions to reach the *pseudo* R^2 quality threshold of 0.95. The *Shepard's diagram* is shown on Figure S3.

Figure 3 shows the low-dimensional projection according to 2 different combinations of *MDS* axes. Figure 3A shows the projection on axes 1 and 2, which reveals a clear separation between the samples of the former (triangles) and later (squares) data acquisitions. This is an obvious batch effect. Using bi-plots to interpret further the direction of separation highlights that this batch effect is almost exclusively explained by a difference in location and scale of the *Live/Dead* channel (Figure S4). This effect is most probably caused by technical differences between the two data acquisitions for this particular channel. These could be, for example, due to different cytometer set-ups or different staining efficiencies for the different batches.

On Figure 3B, samples of the 2 experimental groups (old vs. young donors) can be separated along a direction that is a linear combination of axes 2 and 3. Since we have only included pre-processing channels in the distance calculation, this separation is unlikely to reflect the biological signal of interest. In fact, it probably highlights another batch effect, which is confounded with experimental groups. We show this by interpretation of the direction of separation, which is here less obvious. Bi-plots using medians (Figure S5C), or using standard deviations (not shown), are not able to identify any strong correlation with the direction of separation. However, Figure S5D shows that the latter is strongly correlated with 10th quantiles of *FSC* and *SSC* channels, highlighting the tendency for young adult samples to contain more debris than old adult samples (Figure S5B). This is most probably due to differences in sample handling between these two groups of samples, which were taken from different biobanks (see Methods). Finally, one can also notice that, apart from the first coordinate, driven by *Live/Dead* channel intensities, sample data points with the same label number tend to co-locate (ellipses on Figure 3B). This reflects the fact that these pairs correspond to two acquisitions of the same original biological sample.

To summarize this section, the *CytoMDS* projection of the *ImmunoSenescence Human PBMC* dataset, according to two different combinations of axes, allowed to highlight separations of sample groups that could be attributed to two different batch effects, respectively linked to the data acquisition time point and sample origin. Besides, the proximity of the samples corresponding to the same biological donor could be shown by the projection.

3.3 | Highlighting Biological Signal Between Sample Groups

CytoMDS projections can also reveal biological signals between different groups of samples, provided that this signal is sufficiently strong. In this section, we apply *CytoMDS* on the *Krieg_Anti_PD1* dataset. As this dataset is a mass cytometry dataset, pre-processing involves applying an *arcsinh()* scale transformation, but no channel compensation. Here, we used all

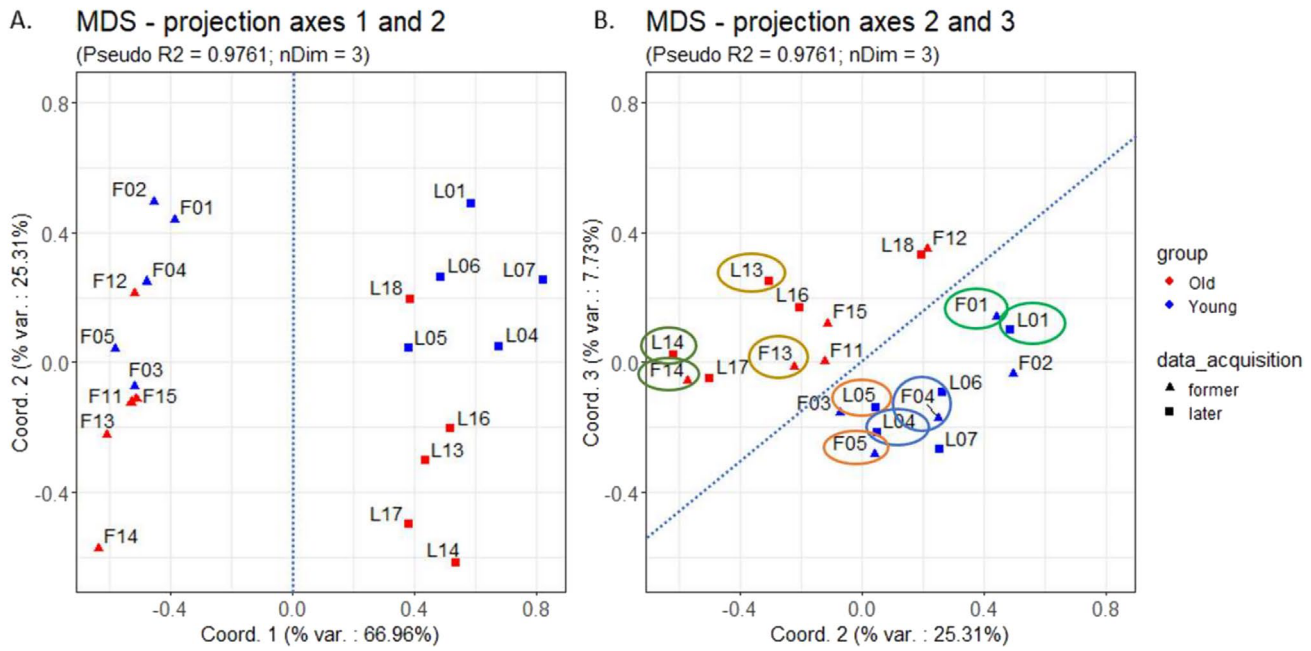

FIGURE 3 | Low dimensional projection of the *ImmunoSenescence Human PBMC* dataset. (A) The projection on axes 1 and 2 reveals a clear batch effect between data points belonging to the former (triangles) and the later data acquisitions (squares). (B) The two groups ('old' in red vs. 'young' in blue), can be separated along a direction that is a linear combination of axes 2 and 3. On the same plot, one can also notice the proximity of the data points corresponding to the biological samples from the same donor (ellipses of different colors: F01 vs. L01, F04 vs. L04, F05 vs. L05, F13 vs. L13, F14 vs. L14). [Color figure can be viewed at wileyonlinelibrary.com]

25 channels as inputs for the *EMD* computation and projected the pairwise distances between the 20 samples using *CytoMDS* with 2 dimensions (Figure 4A).

The *Krieg_Anti_PD1* dataset contains a strong batch effect due to sample acquisition on two different days [28]. This batch effect is clearly visible in Figure 4A: *batch29* samples (squares) tend to be located on the left of the plot, while the *batch23* samples (triangles) tend to be located on the right. Therefore, we performed a batch correction to this dataset by applying the *ComBat* method [32], implemented in the *cyCombine* package [19], and using the responder status—on anti-PD-1 immunotherapy—as a covariate. Figure 4B displays the projection after batch correction, showing a more sensible separation of the two biological condition groups—responders (blue) versus nonresponders (red)—and a better mixing of samples from the two batches.

Figure S6 shows bi-plots with medians for the projection on axes 1 and 2 after batch correction. In the original study [28], it was highlighted that the frequency of a small subpopulation of monocytes in baseline samples was positively associated with responder status following immunotherapy treatment. The left bi-plot of Figure S6 shows that the 11 markers involved in the definition of this rare population tend to be positively correlated with the direction of separation between responders versus non-responders, while this is not the case on the right bi-plot, which shows the other 14 markers.

As a conclusion of this section, we showed that *CytoMDS* can be used as an exploratory data visualization tool to highlight biological signals between groups of samples, provided this signal is sufficiently strong.

4 | Discussion

4.1 | *CytoMDS* Enables Fast and Global Detection of Data Quality Issues

In the Results section, we have provided three examples of using *CytoMDS* to detect sample quality issues. For the *HBV Chronic Mouse* dataset, the issue was specific to a group of samples for which an improper sample handling problem had shrunk the population of interest to a small percentage of the acquired events. For the *ImmunoSenescence Human PBMC* dataset and the *Krieg_Anti_PD1* dataset, the issues were the presence of one or several batch effects, sometimes confounded with the biological condition of interest. Beyond the identification of these quality issues, the ability to link these to specific sample characteristics, hence to potential root causes, is essential to the users, who will more easily make educated decisions about the encountered problems. For this purpose, the *CytoMDS* method has several strengths:

- It generates a global view of all samples, where the distances between points can be quantitatively interpreted.
- It provides bi-plots, a powerful interpretation tool which enables to correlate plot coordinates to sample characteristics. As a benefit, valuable insight on the potential cause(s) of the observed pattern(s) can be inferred.
- It allows unexpected batch effects to be revealed, in addition to those suspected. Indeed, it is not necessary to know in advance which samples belong to which batch. The *MDS* projection itself can be used to uncover patterns, which can in turn be dissected with the help of bi-plots in order to highlight the specific sample characteristics that explain

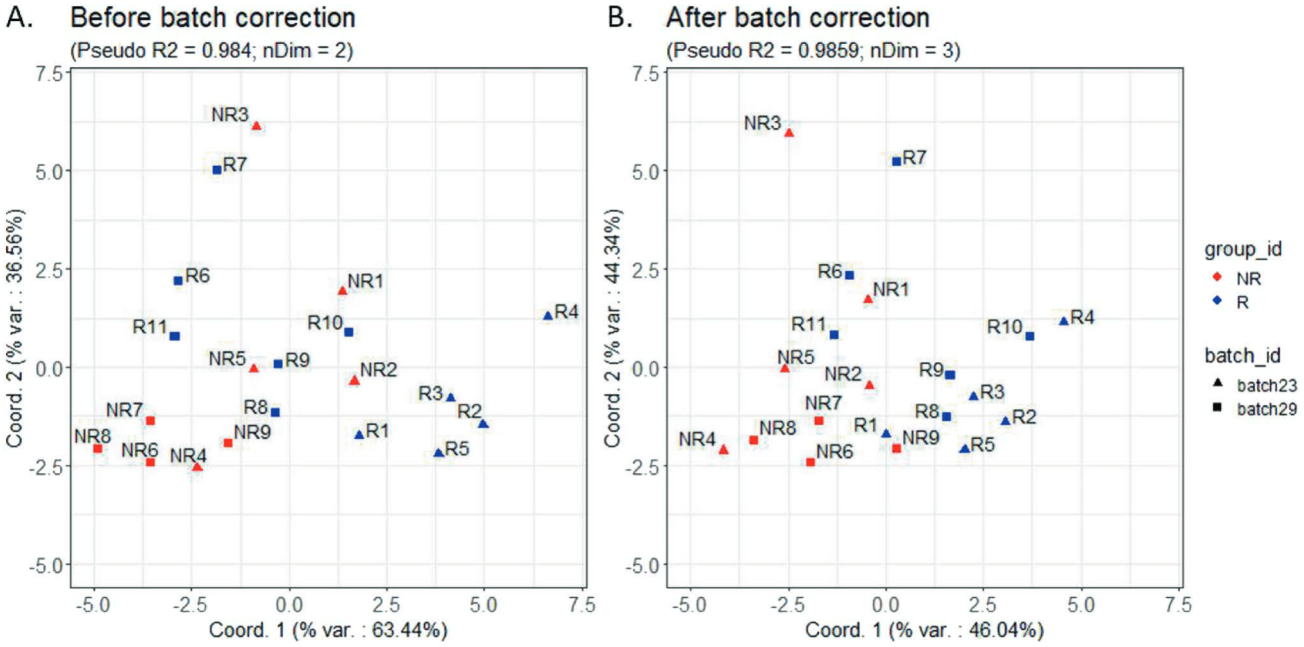

FIGURE 4 | Low dimensional projection of the *Krieg_Anti_PD1* dataset, before batch correction [A], and after batch correction [B], on the first two axes. (A) The projection reveals a clear batch effect between *batch29* samples (squares), which tend to be located on the left side of the plot, and the *batch23* samples (triangles), which tend to be located on the right side. (B) The batch correction results in a better mix between samples of different batches. As a consequence, the separation between the two condition groups—responders (blue) versus non responders (red)—is more sensible. [Color figure can be viewed at wileyonlinelibrary.com]

the observed groupings of samples. Similarly, this process can be used to detect outlier samples.

These strengths are not found in tSNE or UMAP, which are other popular techniques used to highlight batch effects and which typically project the events of all samples of a dataset at once with low dimensions. An example of a 2-dimensional tSNE plot of the *HBV Chronic Mouse* dataset is provided in Figure S7. This plot suggests a batch effect driven by acquisition day, while we know from Figure 2 that this effect is driven by a quality issue observed for only a subset of the samples acquired on a specific day. However, as is shown in Figure S7, due to the high number of samples and the high number of cells, there is no easy way to precisely identify the impacted samples, nor to link the quality issue to specific sample characteristics. In contrast, the *CytoMDS* projection (Figure 2) stays informative when the number of samples increases to tens or even a few hundreds.

Nowicka et al. [33] proposed another *MDS* projection method for QC of cytometry data. This method, however, is not based on distributional distances between samples but uses the medians of each channel to summarize the information contained in each sample. As a result, this method can lead to projections that are very similar to the *CytoMDS* ones only when sample distances are driven mostly by the channel medians, but it is insensitive to other differences between sample distributions. Figure S8 shows a comparative example with low dimensional projections of the *ImmunoSenescence Human PBMC* dataset. *CytoMDS* allows identifying the distributional characteristics that drive the dissimilarity between the samples at hand, which are not necessarily the channel medians.

4.2 | *CytoMDS* Can Be Used at Different Stages of the Data Analysis Pipeline

In Figure 1, we proposed a workflow where only compensation and scale transformation were performed prior to *CytoMDS* projection. However, as a data exploration tool, *CytoMDS* can be run at any stage of the data analysis pipeline. In the Results section, the *CytoMDS* projections were compared before and after batch correction. In general, the iterative use of *CytoMDS* can uncover quality issues at any stage of the data analysis pipeline.

4.3 | Computation Time and Scalability to Larger Datasets

In the Supplementary Results, we present a benchmarking of the *CytoMDS EMD* computation time. As is shown in Figures S9 and S10, the number of markers impacts the *EMD* computation time linearly, while sub-sampling allows to mitigate the effect of a large number of events per sample, with no material impact on precision. As a result, a single *EMD* calculation requires less than one second.

However, for datasets with hundreds of cytometry samples, computing the *EMD* between all sample pairs is a heavy computational task. A first pitfall might arise when loading the whole dataset in memory, which might not be possible due to its size. Another challenge is that calculating a matrix of pairwise distances has a computational complexity of $O(N^2)$, which can lead to long computation time for large datasets. The *CytoMDS* package provides two mechanisms to mitigate these issues. First, to handle datasets of greater size than the available computer RAM, *CytoMDS* provides a distance calculation method that dynamically loads data on the fly when calculating a pairwise sample distance, and releases the corresponding memory afterwards. Second, the user can instruct *CytoMDS* to parallelize the distance calculation process. In that case, the calculation engine takes advantage of the *BiocParallel* package [34] to dispatch the calculation on several workers. The calculation engine then automatically creates balanced worker tasks, corresponding to blocks of the distance matrix, minimizing the overlap between data loaded in memory by different workers. These mechanisms have allowed to decrease the computation time of the full pairwise distance matrix of a dataset of 1400 mass cytometry samples with 43 parameters, from about 7 days to about 1.3 h on a 14 cores user laptop [35].

4.4 | Positioning, Limitation and Possible Future Extension

CytoMDS has been developed as an exploratory visualization method, aiming at facilitating the quality control process for outliers and batch effects detection. As was shown in the *Krieg_Anti_PD1* example, this exploratory visualization can also highlight sample clusters that correspond to different conditions of interest, that is, pointing to the existence of a biological signal. However, identification of such a signal requires a dedicated differential expression/abundance analysis method (e.g., [36, 37]). In particular, a biological signal of interest consisting of the existence (or absence) of a rare immune cell population can translate into very subtle sample distribution differences involving conditional distributions of multiple combined marker intensities. These cannot be captured by *EMD* calculated as the sum of marginal distributions of markers, as is done in *CytoMDS* (see Methods). A possible extension of this work could be to refine the calculation of pairwise sample *EMDs* to incorporate the information on joint marker distributions. This would, however, involve a difficult computational challenge, especially in high-dimensional cytometry datasets. Finally, the current implementation has already shown, throughout the different real dataset examples, its ability to highlight various data quality issues and link them to potential root causes, which is the main goal of our method.

Author Contributions

Philippe Hauchamps: conceptualization, methodology, software, writing – original draft. **Simon Delandre:** writing – original draft, methodology. **Stéphane T. Temmerman:** supervision, writing – review and editing. **Dan Lin:** supervision, methodology, writing – review and editing. **Laurent Gatto:** supervision, conceptualization, methodology, software, writing – review and editing.

Acknowledgments

The authors would like to thank Mehdi Hamrouni and Babak Bayat (GSK) for the provision of the *HBV Chronic Mouse* dataset, as well as Sébastien Baudart and Michaël Ska (GSK) for the data acquisition of the *ImmunoSenescence Human PBMC* dataset.

Conflicts of Interest

Simon Delandre, Stéphane T. Temmerman, and Dan Lin are employees of the GSK group of companies, and report ownership of GSK shares. Stéphane T. Temmerman is listed as inventor on patents owned by the GSK group of companies. Philippe Hauchamps is a student at the de Duve Institute (UCLouvain) and participates in a post graduate studentship program at GSK. The other author declares no conflicts of interest.

Data Availability Statement

Raw cytometry data files are available either on Zenodo (DOI:10.5281/zenodo.10572228) or FlowRepository (ID: FR-FCM-ZYL8). The *ImmunoSenescence Human PBMC* dataset is available upon request. *CytoMDS* package is available on Bioconductor: https://www.bioconductor.org/packages/release/bioc/html/CytoMDS.html. All code needed to reproduce the results presented in the current article is available on the following GitHub repository: https://github.com/UCLouvain-CBIO/2024-CytoMDS-code.

References

1. A. J. Kare, L. Nichols, R. Zermeno, M. N. Raie, S. K. Tumbale, and K. W. Ferrara, "OMIP-095: 40-Color Spectral Flow Cytometry Delineates all Major Leukocyte Populations in Murine Lymphoid Tissues," *Cytometry. Part A* 103, no. 11 (2023): 839–850.

2. M. H. Spitzer and G. P. Nolan, "Mass Cytometry: Single Cells, Many Features," *Cell* 165, no. 4 (2016): 780–791.

3. E. Newell, and Y. Cheng, "Mass Cytometry: Blessed With the Curse of Dimensionality," *Nature Immunology* 17, no. 8 (2016): 890–895.

4. F. J. Hartmann, J. Babdor, P. F. Gherardini, et al., "Comprehensive Immune Monitoring of Clinical Trials to Advance Human Immunotherapy," *Cell Reports* 28, no. 3 (2019): 819–831.

5. T. Schmit, M. Klomp, and M. N. Khan, "The Application of Flow Cytometry for Simultaneous and Multi-Parametric Analysis of Heterogenous Cell Populations in Basic and Clinical Research," *Methods in Molecular Biology* 2223 (2021): 183–200.

6. Y. Saeys, S. Van Gassen, and B. N. Lambrecht, "Computational Flow Cytometry: Helping to Make Sense of High-Dimensional Immunology Data," *Nature Reviews. Immunology* 16, no. 7 (2016): 449–462.

7. J. Staats, A. Divekar, J. P. McCoy, Jr., and H. T. Maecker, "Guidelines for Gating Flow Cytometry Data for Immunological Assays," *Immunophenotyping: Methods and Protocols* 2032 (2019): 81–104.

8. T. Liechti, L. M. Weber, T. M. Ashhurst, et al., "An Updated Guide for the Perplexed: Cytometry in the High-Dimensional Era," *Nature Immunology* 22, no. 10 (2021): 1190–1197.

9. D. Hammill, "CytoExploreR: Interactive Analysis of Cytometry Data. R Package Version 1.1.0," 2021, https://github.com/DillonHammill/CytoExploreR.

10. P. Hauchamps, B. Bayat, S. Delandre, et al., "CytoPipeline and CytoPipelineGUI: A Bioconductor R Package Suite for Building and Visualizing Automated Pre-Processing Pipelines for Flow Cytometry Data," *BMC Bioinformatics* 25, no. 1 (2024): 80.

11. X. Chen, B. Zhang, T. Wang, A. Bonni, and G. Zhao, "Robust Principal Component Analysis for Accurate Outlier Sample Detection in RNA-Seq Data," *BMC Bioinformatics* 21, no. 1 (2020): 269.

12. W. W. B. Goh, W. Wang, and L. Wong, "Why Batch Effects Matter in Omics Data, and How to Avoid Them," *Trends in Biotechnology* 35, no. 6 (2017): 498–507, https://doi.org/10.1016/j.tibtech.2017.02.012.

13. K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, no. 11 (1901): 559–572, https://doi.org/10.1080/14786440109462720.

14. W. W. B. Goh and L. Wong, "Protein Complex-Based Analysis Is Resistant to the Obfuscating Consequences of Batch Effects — A Case Study in Clinical Proteomics," *BMC Genomics* 18, no. 2 (2017): 142.

15. P. Hauchamps and L. Gatto, "CytoMDS: Low Dimensions Projection of Cytometry Samples. R Package Version 1.0.0," 2024, https://uclouvain-cbio.github.io/CytoMDS.

16. W. Huber, V. J. Carey, R. Gentleman, et al., "Orchestrating High-Throughput Genomic Analysis With Bioconductor," *Nature Methods* 12, no. 2 (2015): 115–121.

17. Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision* 40, no. 2 (2000): 99–121.

18. D. Y. Orlova, N. Zimmerman, S. Meehan, et al., "Earth Mover's Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell Populations," *PLoS One* 11, no. 3 (2016): e0151859.

19. C. B. Pedersen, S. H. Dam, M. B. Barnkob, et al., "cyCombine Allows for Robust Integration of Single-Cell Cytometry Datasets Within and Across Technologies," *Nature Communications* 13, no. 1 (2022): 1698.

20. S. Van Gassen, B. Gaudilliere, M. S. Angst, Y. Saeys, and N. Aghaeepour, "CytoNorm: A Normalization Algorithm for Cytometry Data," *Cytometry. Part A* 97, no. 3 (2020): 268–278.

21. H. Yi and N. Stanley, "CytoEMD: Detecting and Visualizing Between-Sample Variation in Relation to Phenotype With Earth Mover's Distance," in *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (Association for Computing Machinery, 2022), 1–14.

22. D. Schuhmacher, B. Bähre, C. Gottschlich, V. Hartmann, F. Heinemann, and B. Schmitzer, "Transport: Computation of Optimal Transport Plans and Wasserstein Distances. R Package Version 0.14-7," 2024, https://cran.r-project.org/package=transport.

23. J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction* (Springer, 2007), https://doi.org/10.1007/978-0-387-39351-3.

24. L. Van Der Maaten and G. Hinton, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research* 9, no. 86 (2008): 2579–2605, http://jmlr.org/papers/v9/vandermaaten08a.html.

25. L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv Preprint arXiv:1802.03426* (2018).

26. J. De Leeuw and P. Mair, "Multidimensional Scaling Using Majorization: SMACOF in R," *Journal of Statistical Software* 31 (2009): 1–30.

27. I. Borg, P. J. F. Groenen, and P. Mair, *Applied Multidimensional Scaling and Unfolding* (Springer International Publishing, 2018), https://doi.org/10.1007/978-3-319-73471-2.

28. C. Krieg, M. Nowicka, S. Guglietta, et al., "High-Dimensional Single-Cell Analysis Predicts Response to Anti-PD-1 Immunotherapy," *Nature Medicine* 24, no. 2 (2018): 144–153.

29. L. M. Weber and C. Soneson, "HDCytoData: Collection of High-Dimensional Cytometry Benchmark Datasets in Bioconductor Object Formats," *F1000Res* 8 (2019): 1459.

30. D. R. Parks, M. Roederer, and W. A. Moore, "A New "Logicle" Display Method Avoids Deceptive Effects of Logarithmic Scaling for Low Signals and Compensated Data," *Cytometry. Part A* 69, no. 6 (2006): 541–551.

31. H. Den Braanker, M. Bongenaar, and E. Lubberts, "How to Prepare Spectral Flow Cytometry Datasets for High Dimensional Data Analysis: A Practical Workflow," *Frontiers in Immunology* 12 (2021): 768113.

32. W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods," *Biostatistics* 8, no. 1 (2007): 118–127.

33. M. Nowicka, C. Krieg, H. L. Crowell, et al., "CyTOF Workflow: Differential Discovery in High-Throughput High-Dimensional Cytometry Datasets," *F1000Res* 6 (2017): 748.

34. M. Morgan, J. Wang, V. Obenchain, M. Lang, R. Thompson, and N. Turaga, "BiocParallel: Bioconductor Facilities for Parallel Evaluation. R Package Version 1.37.0," (2023), https://bioconductor.org/packages/BiocParallel.

35. S. Granjeaud, N. Abdellaoui, A.-S. Chretien, et al., "CytoBatchNorm: An R Package With Graphical Interface for Batch Effects Correction of Cytometry Data," *bioRxiv* (2024), https://doi.org/10.1101/2024.05.29.596492.

36. C. Seiler, A.-M. Ferreira, L. M. Kronstad, et al., "CytoGLMM: Conditional Differential Analysis for Flow and Mass Cytometry Experiments," *BMC Bioinformatics* 22, no. 1 (2021): 137.

37. L. M. Weber, M. Nowicka, C. Soneson, and M. D. Robinson, "Diffcyt: Differential Discovery in High-Dimensional Cytometry via High-Resolution Clustering," *Communications Biology* 2 (2019): 183, https://doi.org/10.1038/s42003-019-0415-5.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.