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, Belgium; [2]GSK, Rixensart, Belgium.

✉ **For correspondence:** laurent.gatto@uclouvain.be

Present address: Computational Biology and Bioinformatics, de Duve Institute, UCLouvain, Av. Hippocrate 75, 1200 Brussels, Belgium

Data availability: Raw cytometry data files are available either on Zenodo (DOI:10.5281/zenodo.8425840) or FlowRepository (ID: FR-FCM-ZYL8). The *ImmunoSenescence Human PBMC* dataset will be made soon available, upon request.

Code availability: All code needed to reproduce the results presented in the current article is available on the following GitHub repository: *github.com/UCLouvain-CBIO/2024-CytoMDS-code*.

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

Competing interests: S.D., S.T., and D.L. are employees of the GSK group of companies, and report ownership of GSK shares. S.T. is listed as inventor on patents owned by the GSK group of companies. P.H. is a student at the de Duve Institute (UCLouvain) and participates in a post graduate studentship program at GSK.

Abstract

Quality Control (QC) of samples is an essential preliminary step in cytometry data analysis. Notably, identification of potential batch effects and outlying samples is paramount, to avoid mistaking these effects for true biological signal in downstream analyses. However, this task can prove to be delicate and tedious, especially for datasets with many 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. It 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.

Introduction

Recent technology advances in flow and mass cytometry allow the acquisition of high dimensional sample data, with up to respectively 40 and 50 parameters measured simultaneously for millions of cells (Kare et al., 2023, Spitzer and Nolan, 2016). For such high dimensional data samples, extracting relevant biological information becomes challenging (Newell and Cheng, 2016). Moreover, in large clinical studies, possibly with some longitudinal and multicentric components, this difficult task is itself multiplied by the high number of samples (Hartmann et al., 2019, Schmit, Klomp, and Khan, 2021). In the last decade, this led to the development of (semi) automated data analysis pipelines (Saeys, Van Gassen, and Lambrecht, 2016), to complement the traditional manual gating

approaches (Staats et al., 2019). 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 (Liechti et al., 2021): compensation, stability of signal in time, removal of undesirable events like debris, doublets or dead cells, batch effect detection and removal, etc.

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 e.g. FlowJo or Cytobank - among the available commercial platforms, or free open source solutions (Hammill, 2021, Hauchamps et al., 2024). 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's, namely outlying sample identification (Chen et al., 2020) and batch effect detection (Goh, Wang, and Wong, 2017), would benefit from a global representation of the dataset. Well-known classical data exploration techniques like *Principal Component Analysis (PCA)* (Pearson, 1901) have the potential to provide such a global view of the data, and are routinely used in other omics context like e.g. bulk transcriptomics (Chen et al., 2020) and bulk proteomics (Goh and Wong, 2017) data analyses. Applying these techniques to cytometry data nonetheless presents the challenge that, in single cell data, each sample consists of a data matrix, which does not fit the use of a standard *PCA* or a similar algorithm.

To address this challenge, we developed *CytoMDS* (Hauchamps and Gatto, 2024), 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 (Huber et al., 2015), as of version 3.19.

Materials and Methods

Proposed method

Figure 1 sketches an overview of the *CytoMDS* workflow. As input, the method requires a set of minimally pre-processed - i.e. scale transformed and, if applicable, compensated or unmixed - sample files in *flow cytometry standard (fcs)* format (Figure 1, [A]). The *CytoMDS* method itself (Figure 1, [B]) 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).

Optionally, *bi-plots* can be overlaid on the projection. These consist in visual representations of correlations (4) of chosen sample statistics (3) with the projection coordinates, for interpreting directions within the plot.

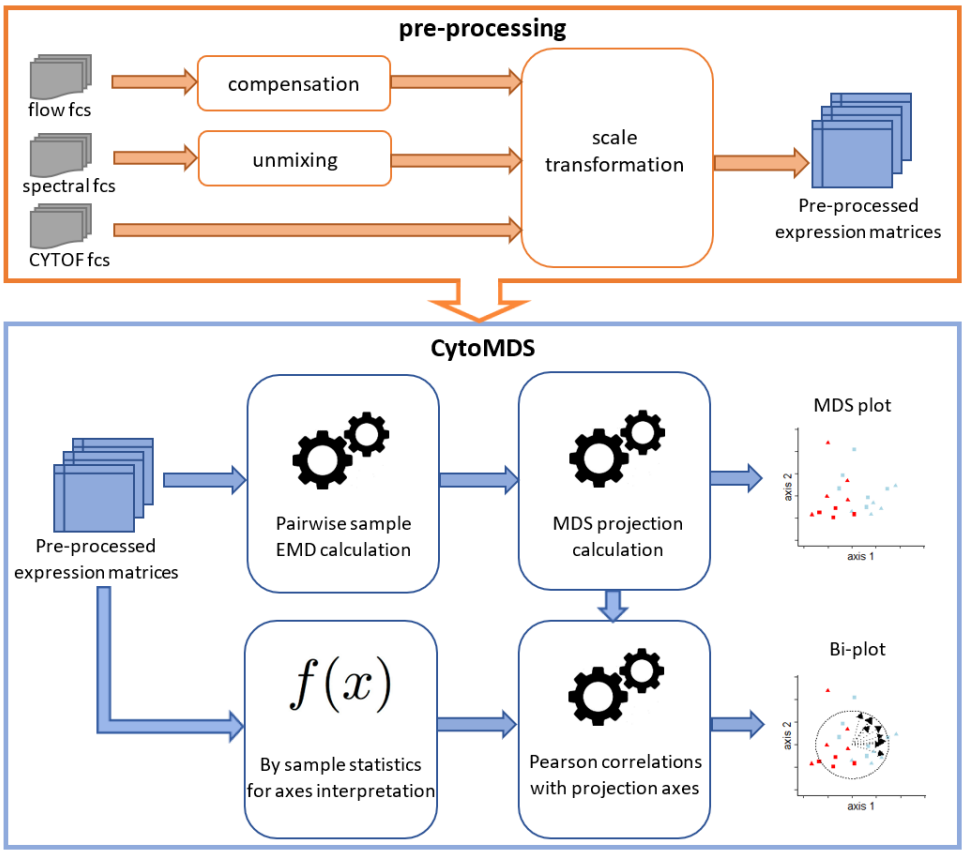

Figure 1. Overview of the *CytoMDS* workflow. Pre-processing of the raw *fcs* files [A] should be done according to the workflow in the orange box, which includes scale transformation, and might include compensation or unmixing depending on the type of data. The *CytoMDS* method itself [B] takes as input 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 a *MDS* plot. Optionally, in order to allow interpretation of *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).

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 distribution 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 (Rubner, Tomasi, and Guibas, 2000).

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 (Orlova et al., 2016). It has been used in different applications of cytometry data analysis, especially in the evaluation of batch correction methods (Van Gassen et al., 2020, Pedersen et al., 2022) or for obtaining a sample representation in relation with different phenotypes (Yi and Stanley, 2022).

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 $m : 1,\ldots M$. 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 $(i : \{1, \dots r\})$ 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 $(j : \{1, \dots s\})$ 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: $F(t) = \sum_{i=1}^{r} w_i^a 1\{a_i \leq t\}$ and $G(t) = \sum_{j=1}^{s} w_j^b 1\{b_j \leq t\}$.
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* (Schuhmacher et al., 2024). This process is then repeated for each pair of samples, leading to a symmetric pairwise distance matrix having *N* rows and *N* columns - with *N* the number of samples. Its diagonal is filled with zeros.

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 (Lee and Verleysen, 2007), in contrast to popular local neighbourhood preserving methods such as *t-SNE* (van der Maaten and Hinton, 2008) and *UMAP* (McInnes, Healy, and Melville, 2018).
Furthermore, since the *EMD* is a metric but is not euclidean, we need to resort, for the projection, to *non classical metric MDS*. The latter is a non linear method minimizing a specific criterion called *stress* (Lee and Verleysen, 2007), which represent a quadratic error of projection and is defined as follows:

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

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 (de Leeuw and Mair, 2009). 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.

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 Figure S1 and Figure 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:

$$pseudoR^2 = 1 - \frac{\sum_{i<j} (\delta_{ij} - d_{ij}(\boldsymbol{X}))^2}{\sum_{i<j} (\delta_{ij} - \bar{\delta})^2} \text{ , where } \bar{\delta} \text{ is the overall mean of pairwise } EMD. \tag{3}$$

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, note that additional indicators, output of the *SMACOF* algorithm, such as the minimized value for the stress criterion (Equation 2), and the contribution of each sample to the stress, i.e. *stress per point* (de Leeuw and Mair, 2009), can also be retrieved.

Coordinates and directions interpretation using bi-plots

Bi-plots are plots that are overlaid to the *MDS* projection, allowing to interpret directions of interest by relating them to well-chosen sample characteristics (Borg, Groenen, and Mair, 2018). Creating a bi-plot involves choosing *external scales*, i.e. sample specific characteristics that can be fitted to the *MDS* space. Possible choices are summary statistics of marker intensities, e.g. univariate mean, standard deviation, median or other quantiles, etc., 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 overlay 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 landmark for visual convenience (e.g. in Figure 2, [C] and [D]).

Example datasets

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

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 regimes on the immune response of Hepatitis B Virus transduced mice. This dataset has been described in Hauchamps et al., 2024. In short, it contains 55 flow cytometry mouse liver lymphocyte samples, analyzed with a 12-channel flow cytometry panel. The 55 mice were split in 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).

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 markers 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 ageing - 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.

Krieg_Anti_PD1 dataset

The *Krieg_Anti_PD1* dataset is a mass cytometry dataset (Krieg et al., 2018), 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 in 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* (Weber and Soneson, 2019).

Results

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 bi-exponential transformations (Parks, Roederer, and Moore, 2006). *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 of the *EMD* calculation, i.e. *FSC-A*, *FSC-H*, *SSC-A* and the *Live/Dead* channels, which are the channels used for pre-gating (den Braanker, Bongenaar, and Lubberts, 2021). 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 degree identity line. Also, the *pseudo* R^2 is 0.9679, hence above the 0.95 threshold (chosen by default), with only two projected dimensions.

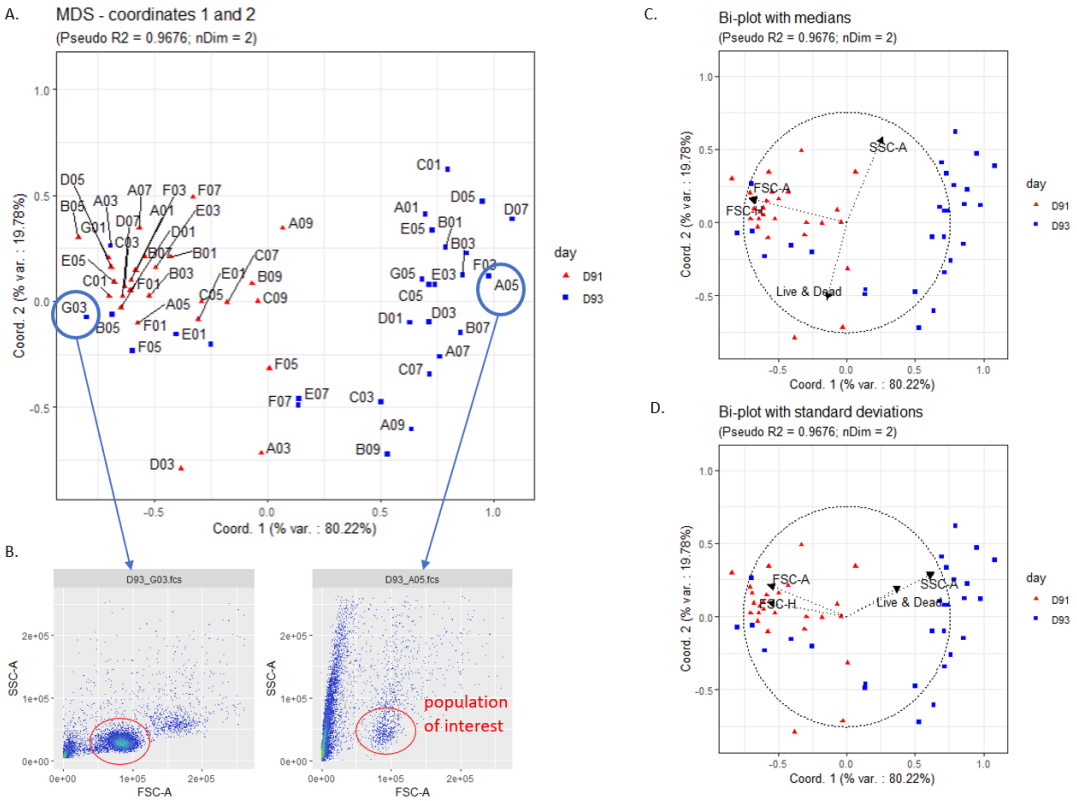

Figure 2. The low dimensional projection of the *HBV Chronic Mouse* dataset [A] reveals a clear split between two sample groups, not fully attributable to a batch effect linked to day of acquisition - since a minority of day D93 samples belong to the left sample group. Bi-plots overlaying respectively 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* vs. *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.

The projection, in Figure 2 [A], 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, 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 2 [B], showing 2D plots for two representative samples, selected for their extreme opposite position on the first axis. 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 to identify a sample cluster corresponding to low quality samples, in terms of relative percentage of events belonging to the cell population of interest.

Highlighting batch effects

In this section, we apply *CytoMDS* on the *ImmunoSenescence Human PBMC* dataset. As in the previous section, we pre-processed the data by applying compensation - based on a manually adjusted compensation matrix - and scale transformed the 26 channel intensities. 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 combination of *MDS* axes. [A] 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 of location and scale of the *Live/Dead* channel (Figure S4). This effect is most probably caused by technical differences between the two data acquisition 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 3 [B], 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 S5 [C]), or using standard deviations (not shown), are not able to identify any strong correlation with the direction of separation. However, Figure S5 [D] 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 S5 [B]). 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 colocate (ellipses on Figure 3 [B]). 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 data acquisition time

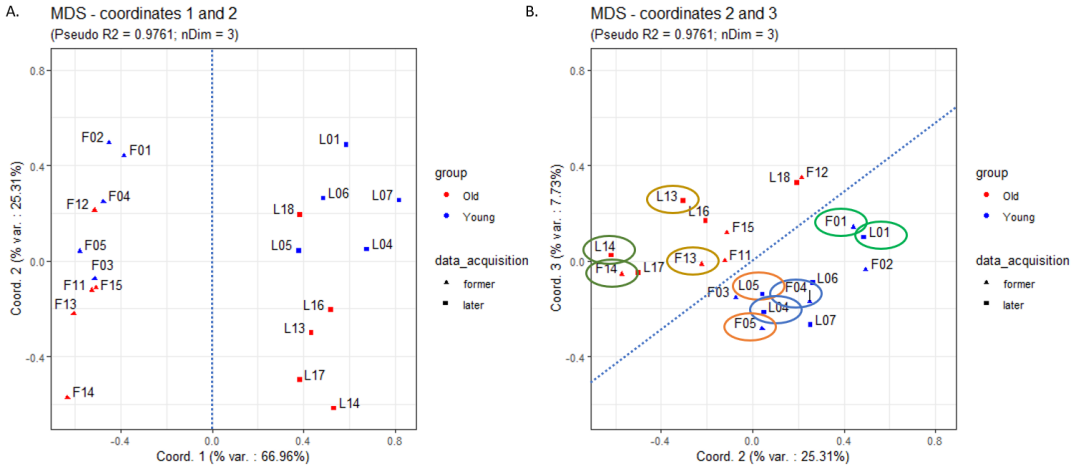

Figure 3. Low dimensional projection of the *ImmunoSenescence Human PBMC* dataset. (A) The projection on axes 1 and 2 reveals a clear batch effect between data points belonging to the former (triangles) and the later data acquisitions (squares). (B) The two groups ('old' in red vs. 'young' in blue), 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 colours: F01 vs. L01, F04 vs. L04, F05 vs. L05, F13 vs. L13, F14 vs. L14).

point and sample origin. Besides, proximity of the samples corresponding to the same biological donor could be shown by the projection.

Highlighting biological signal between sample groups

CytoMDS projections can also reveal biological signal 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 25 channels as input for the *EMD* computation, and projected the pairwise distances between the 20 samples using *CytoMDS* with 2 dimensions (Figure 4, [A]).

The *Krieg_Anti_PD1* dataset contains a strong batch effect, due to sample acquisition on two different days (Krieg et al., 2018). This batch effect is clearly visible on Figure 4, [A]: *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 (Johnson, Li, and Rabinovic, 2007), implemented in the *cyCombine* package (Pedersen et al., 2022), and using the responder status - on anti-PD-1 immunotherapy - as covariate. Figure4 [B] displays the projection after batch correction, showing a more sensible separation of the two biological condition groups - responders (blue) vs. non responders (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 (Krieg et al., 2018), 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 vs 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 signal between groups of samples, provided this signal is sufficiently strong.

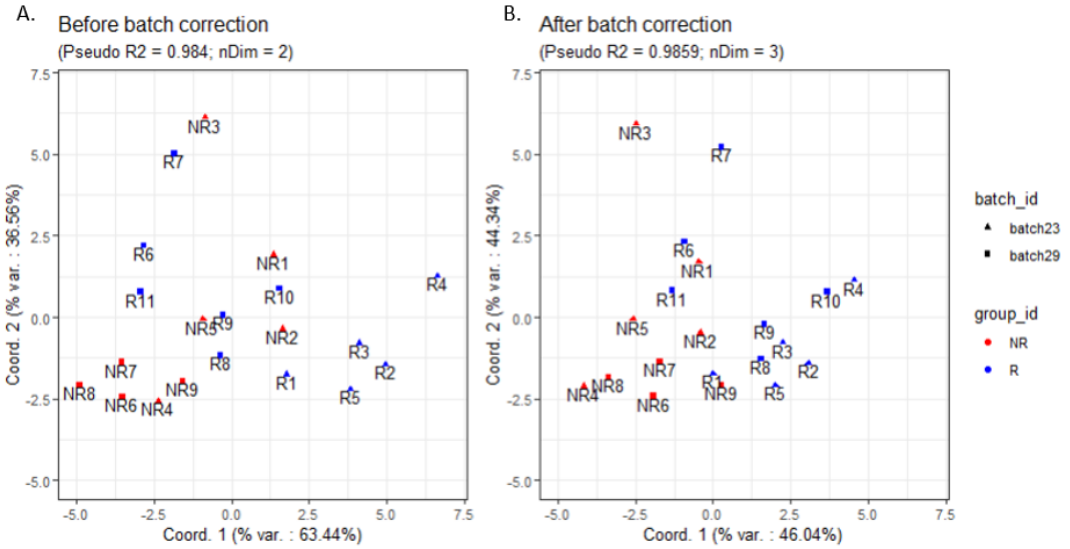

Figure 4. Low dimensional projection of the *Krieg_Anti_PD1* dataset, before batch correction [A], and after batch correction [B], on the first two axes. [A] The projection reveals a clear batch effect between *batch29* samples (squares), which tend to be located on the left side of the plot, and the *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) vs. non responders (red) - is more sensible.

Discussion

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.
- When it comes to batch effect identification, knowing in advance which samples belong to which batch is not required. The *MDS* projection itself can be used to uncover patterns, which in turn can be further dissected with the help of bi-plots, to highlight specific sample characteristics that explain observed sample clusters. This process allows to reveal unexpected, on top of suspected, batch effects. The same applies to detection of outlying 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 sam-

ples acquired on a specific day. However, as is shown in Figure S7, due to the high number of samples and 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., 2017 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 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 to identify the distributional charcateristics that drive the dissimilarity between the samples at hand, which are not necessarily the channel medians.

Scalability to larger datasets

Computing *EMD* between all sample pairs for large datasets (i.e. with hundreds of samples), 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 (Morgan et al., 2023) 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 hours on a 14 cores user laptop (Granjeaud et al., 2024).

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, i.e. pointing to the existence of a biological signal. However, identification of such signal requires a dedicated differential expression/abundance analysis method (e.g. Weber et al., 2019, Seiler et al., 2021). In particular, a biological signal of interest consisting in 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 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 involves a difficult computational challenge, especially in high dimensional cytometry datasets. Besides, 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.

Declarations

Funding

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

Competing Interests

- S.D., S.T., and D.L. are employees of the GSK group of companies and report ownership of GSK shares.
- S.T. is listed as inventor on patents owned by the GSK group of companies.
- P.H. is a student at the de Duve Institute (UCLouvain) and participates in a post graduate studentship program at GSK.
- L.G. reports no competing interest.

Author contributions

P.H.: conceptualization (lead); methodology; software (lead); writing - original draft (lead). S.D.: methodology (supporting); writing - original draft (supporting). S.T.: supervision (supporting); writing - review & editing (supporting). D.L.: supervision; methodology (supporting); writing - review & editing. L.G.: supervision; conceptualization (supporting); methodology; software (supporting); writing - review & editing.

Availability of data and materials

Raw cytometry data files are available either on Zenodo (DOI:10.5281/zenodo.8425840) or FlowRepository (ID: FR-FCM-ZYL8). The *ImmunoSenescence Human PBMC* dataset will be made soon available, upon request.

Code availability

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.

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.

This preprint was created using the LaPreprint template (https://github.com/roaldarbol/lapreprint) by Mikkel Roald-Arbøl.

References

Borg, Ingwer, Patrick J.F. Groenen, and Patrick Mair (2018). *Applied Multidimensional Scaling and Unfolding*. Springer International Publishing. ISBN: 9783319734712. DOI: 10.1007/978-3-319-73471-2. URL: http://dx.doi.org/10.1007/978-3-319-73471-2.

Chen, Xiaoying, Bo Zhang, Ting Wang, Azad Bonni, and Guoyan Zhao (2020). "Robust principal component analysis for accurate outlier sample detection in RNA-Seq data". In: *BMC Bioinformatics* 21.1, p. 269.

de Leeuw, Jan and Patrick Mair (2009). "Multidimensional Scaling Using Majorization: SMACOF in R". In: *J. Stat. Softw.* 31, pp. 1–30.

den Braanker, Hannah, Margot Bongenaar, and Erik Lubberts (2021). "How to Prepare Spectral Flow Cytometry Datasets for High Dimensional Data Analysis: A Practical Workflow". In: *Front. Immunol.* 12, p. 768113.

Goh, Wilson Wen Bin, Wei Wang, and Limsoon Wong (2017). "Why Batch Effects Matter in Omics Data, and How to Avoid Them". In: *Trends Biotechnol.* 35.6, pp. 498–507.

Goh, Wilson Wen Bin and Limsoon Wong (2017). "Protein complex-based analysis is resistant to the obfuscating consequences of batch effects — a case study in clinical proteomics". In: *BMC Genomics* 18.Suppl 2, p. 142.

Granjeaud, Samuel, Naoill Abdellaoui, Anne-Sophie Chretien, Eloise Woitrain, Laurent Pineau, Sandro Ninni, Alexandre Harari, Marion Arnaud, David Montaigne, Bart Staels, David Dombrowicz, and Olivier Molendi-Coste (2024). "CytoBatchNorm: an R package with graphical interface for batch effects correction of cytometry data".

Hammill, Dillon (2021). *CytoExploreR: Interactive Analysis of Cytometry Data*. R package version 1.1.0. URL: https://github.com/DillonHammill/CytoExploreR.

Hartmann, Felix J, Joel Babdor, Pier Federico Gherardini, El-Ad D Amir, Kyle Jones, Bita Sahaf, Diana M Marquez, Peter Krutzik, Erika O'Donnell, Natalia Sigal, Holden T Maecker, Everett Meyer, Matthew H Spitzer, and Sean C Bendall (2019). "Comprehensive Immune Monitoring of Clinical Trials to Advance Human Immunotherapy". In: *Cell Rep.* 28.3, 819–831.e4.

Hauchamps, Philippe, Babak Bayat, Simon Delandre, Mehdi Hamrouni, Marie Toussaint, Stephane Temmerman, Dan Lin, and Laurent Gatto (2024). "CytoPipeline and CytoPipelineGUI: a Bioconductor R package suite for building and visualizing automated pre-processing pipelines for flow cytometry data". In: *BMC Bioinformatics* 25.1, p. 80.

Hauchamps, Philippe and Laurent Gatto (2024). *CytoMDS: Low Dimensions projection of cytometry samples*. R package version 1.0.0. URL: https://uclouvain-cbio.github.io/CytoMDS.

Huber, Wolfgang, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D Hansen, Rafael A Irizarry, Michael Lawrence, Michael I Love, James MacDonald, Valerie Obenchain, Andrzej K Oleś, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K Smyth, Dan Tenenbaum, Levi Waldron, and Martin Morgan (2015). "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nat. Methods* 12.2, pp. 115–121.

Johnson, W Evan, Cheng Li, and Ariel Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1, pp. 118–127.

Kare, Aris J, Lisa Nichols, Ricardo Zermeno, Marina N Raie, Spencer K Tumbale, and Katherine W Ferrara (2023). "OMIP-095: 40-Color spectral flow cytometry delineates all major leukocyte populations in murine lymphoid tissues". In: *Cytometry A* 103.11, pp. 839–850.

Krieg, Carsten, Malgorzata Nowicka, Silvia Guglietta, Sabrina Schindler, Felix J Hartmann, Lukas M Weber, Reinhard Dummer, Mark D Robinson, Mitchell P Levesque, and Burkhard Becher (2018). "High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy". In: *Nat. Med.* 24.2, pp. 144–153.

Lee, John A and Michel Verleysen (2007). *Nonlinear Dimensionality Reduction*. Springer New York. ISBN: 9780387393506. DOI: 10.1007/978-0-387-39351-3. URL: http://dx.doi.org/10.1007/978-0-387-39351-3.

Liechti, Thomas, Lukas M Weber, Thomas M Ashhurst, Natalie Stanley, Martin Prlic, Sofie Van Gassen, and Florian Mair (2021). "An updated guide for the perplexed: cytometry in the high-dimensional era". In: *Nat. Immunol.* 22.10, pp. 1190–1197.

McInnes, Leland, John Healy, and James Melville (2018). "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: arXiv: 1802.03426 [stat.ML].

Morgan, Martin, Jiefei Wang, Valerie Obenchain, Michel Lang, Ryan Thompson, and Nitesh Turaga (2023). *BiocParallel: Bioconductor facilities for parallel evaluation*. R package version 1.37.0. DOI: 10.18129/B9.bioc.BiocParallel. URL: https://bioconductor.org/packages/BiocParallel.

Newell, Evan W and Yang Cheng (2016). "Mass cytometry: blessed with the curse of dimensionality". In: *Nat. Immunol.* 17.8, pp. 890–895.

Nowicka, Malgorzata, Carsten Krieg, Helena L Crowell, Lukas M Weber, Felix J Hartmann, Silvia Guglietta, Burkhard Becher, Mitchell P Levesque, and Mark D Robinson (2017). "CyTOF workflow:

differential discovery in high-throughput high-dimensional cytometry datasets". In: *F1000Res.* 6, p. 748.

Orlova, Darya Y, Noah Zimmerman, Stephen Meehan, Connor Meehan, Jeffrey Waters, Eliver E B Ghosn, Alexander Filatenkov, Gleb A Kolyagin, Yael Gernez, Shanel Tsuda, Wayne Moore, Richard B Moss, Leonore A Herzenberg, and Guenther Walther (2016). "Earth Mover's Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell Populations". In: *PLOS ONE* 11.3, e0151859.

Parks, David R, Mario Roederer, and Wayne A Moore (2006). "A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data". In: *Cytometry A* 69.6, pp. 541–551.

Pearson, Karl (1901). "On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, pp. 559–572. DOI: 10.1080/14786440109462720.

Pedersen, Christina Bligaard, Søren Helweg Dam, Mike Bogetofte Barnkob, Michael D Leipold, Noelia Purroy, Laura Z Rassenti, Thomas J Kipps, Jennifer Nguyen, James Arthur Lederer, Satyen Harish Gohil, Catherine J Wu, and Lars Rønn Olsen (2022). "cyCombine allows for robust integration of single-cell cytometry datasets within and across technologies". In: *Nat. Commun.* 13.1, p. 1698.

Rubner, Yossi, Carlo Tomasi, and Leonidas J Guibas (2000). "The Earth Mover's Distance as a Metric for Image Retrieval". In: *Int. J. Comput. Vis.* 40.2, pp. 99–121.

Saeys, Yvan, Sofie Van Gassen, and Bart N Lambrecht (2016). "Computational flow cytometry: helping to make sense of high-dimensional immunology data". In: *Nat. Rev. Immunol.* 16.7, pp. 449–462.

Schmit, Taylor, Mitchell Klomp, and M Nadeem Khan (2021). "The Application of Flow Cytometry for Simultaneous and Multi-parametric Analysis of Heterogenous Cell Populations in Basic and Clinical Research". In: *Methods Mol. Biol.* 2223, pp. 183–200.

Schuhmacher, Dominic, Björn Bähre, Carsten Gottschlich, Valentin Hartmann, Florian Heinemann, and Bernhard Schmitzer (2024). *transport: Computation of Optimal Transport Plans and Wasserstein Distances*. R package version 0.14-7. URL: https://cran.r-project.org/package=transport.

Seiler, Christof, Anne-Maud Ferreira, Lisa M Kronstad, Laura J Simpson, Mathieu Le Gars, Elena Vendrame, Catherine A Blish, and Susan Holmes (2021). "CytoGLMM: conditional differential analysis for flow and mass cytometry experiments". In: *BMC Bioinformatics* 22.1, p. 137.

Spitzer, Matthew H and Garry P Nolan (2016). "Mass Cytometry: Single Cells, Many Features". In: *Cell* 165.4, pp. 780–791.

Staats, Janet, Anagha Divekar, J Philip McCoy Jr, and Holden T Maecker (2019). "Guidelines for Gating Flow Cytometry Data for Immunological Assays". In: *Methods Mol. Biol.* 2032, pp. 81–104.

van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Van Gassen, Sofie, Brice Gaudilliere, Martin S Angst, Yvan Saeys, and Nima Aghaeepour (2020). "CytoNorm: A Normalization Algorithm for Cytometry Data". In: *Cytometry A* 97.3, pp. 268–278.

Weber, Lukas M, Malgorzata Nowicka, Charlotte Soneson, and Mark D Robinson (2019). "diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering". In: *Commun Biol* 2, p. 183.

Weber, Lukas M and Charlotte Soneson (2019). "HDCytoData: Collection of high-dimensional cytometry benchmark datasets in Bioconductor object formats". In: *F1000Res.* 8, p. 1459.

Yi, Haidong and Natalie Stanley (2022). "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*. BCB '22 Article 28. Northbrook, Illinois: Association for Computing Machinery, pp. 1–14.

Supplementary Materials

Supplementary Tables

Donor ID	Group	Former data acquisition	Later data acquisition
1	Younger Adult	F1	L1
2	Younger Adult		L2
3	Younger Adult		L3
4	Younger Adult	F4	L4
5	Younger Adult	F5	L5
6	Younger Adult	F6	
7	Younger Adult	F7	
11	Older Adult		L11
12	Older Adult		L12
13	Older Adult	F13	L13
14	Older Adult	F14	L14
15	Older Adult		L15
16	Older Adult	F16	
17	Older Adult	F17	
18	Older Adult	F18	

Table S1. Sample groups and data acquisitions for the *ImmunoSenescence Human PBMC* dataset.

Supplementary Figures

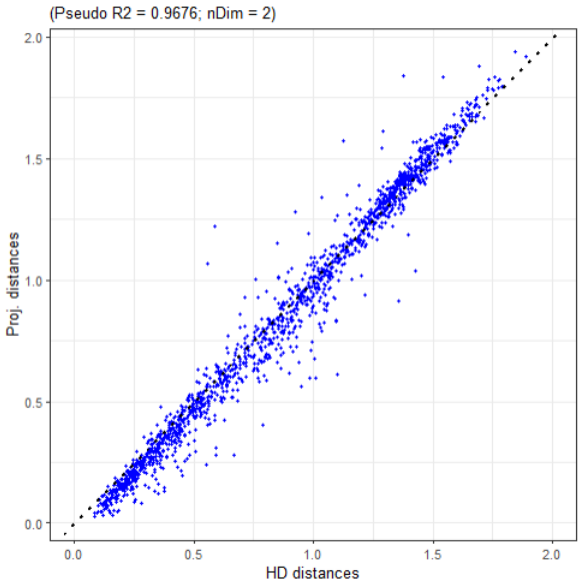

Figure S1. *Shepard's diagram* for the projection of the *HBV Chronic Mouse* dataset. Each dot represents a sample pair, where the x value is the high dimensional *Earth Mover's Distance* between the two samples and the y value is the obtained euclidian distance on the low dimensional projection. A projection of high quality results in most points being close to the ideal 45 degree identity line.

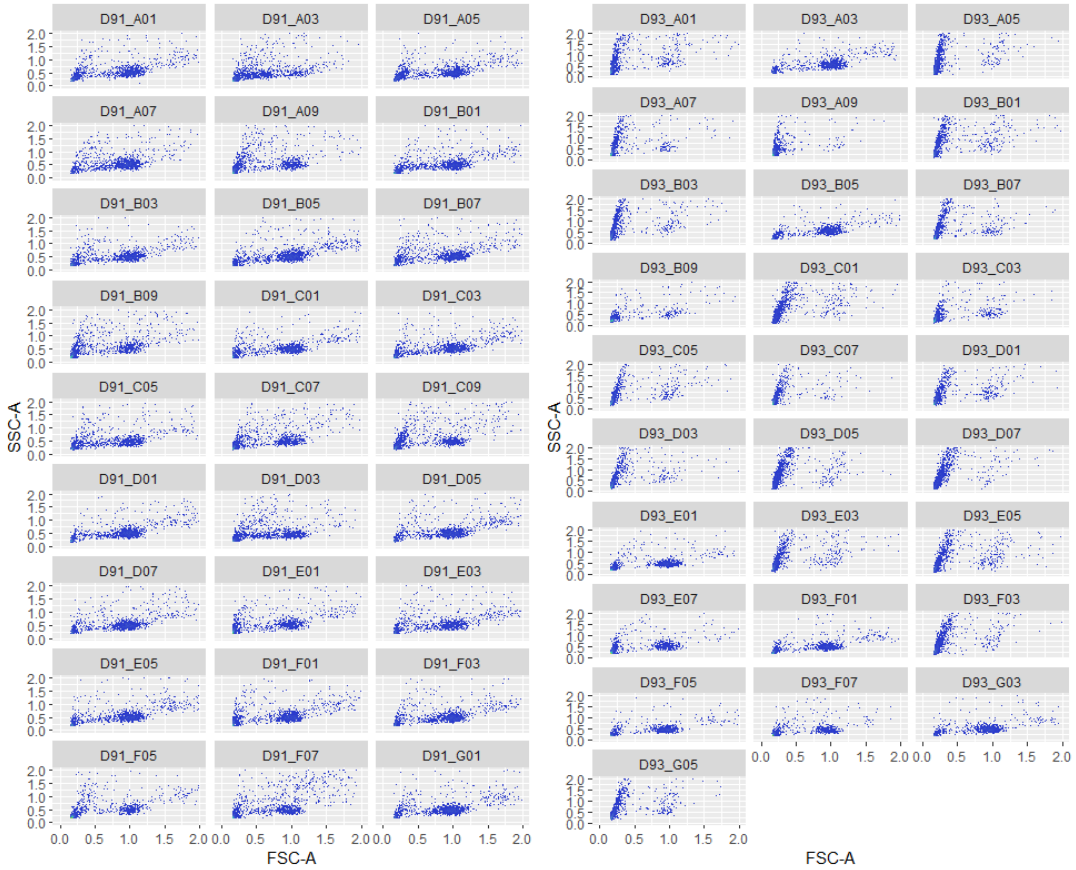

Figure S2. *FSC-A* vs. *SSC-A* 2D visualization of the 55 samples of the *HBV Chronic Mouse* dataset, revealing a clear difference of pattern between low quality samples and good quality samples, corresponding to the two sample groups in Figure 2. Compared to Figure 2, here the axes have undergone a linear mapping to the [0.0, 2.0] range, in order to keep the axis legends readable.

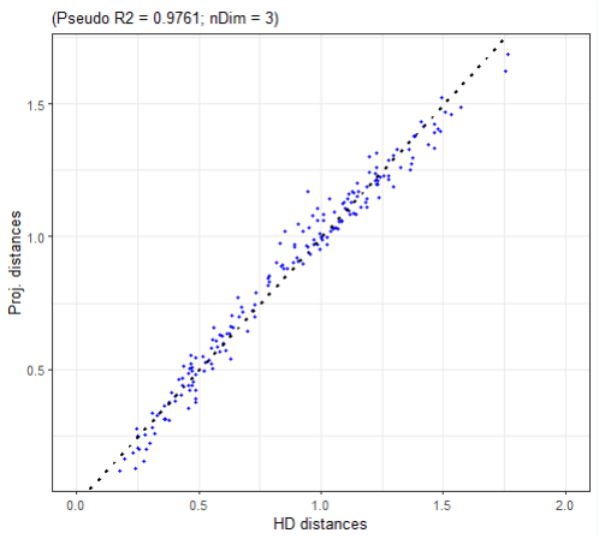

Figure S3. *Shepard's diagram* for the projection of the *ImmunoSenescence Human PBMC* dataset. Each dot represents a sample pair, where the x value is the high dimensional *Earth Mover's Distance* between the two samples and the y value is the obtained euclidian distance on the low dimensional projection. A projection of high quality results in most points being close to the ideal 45 degree identity line.

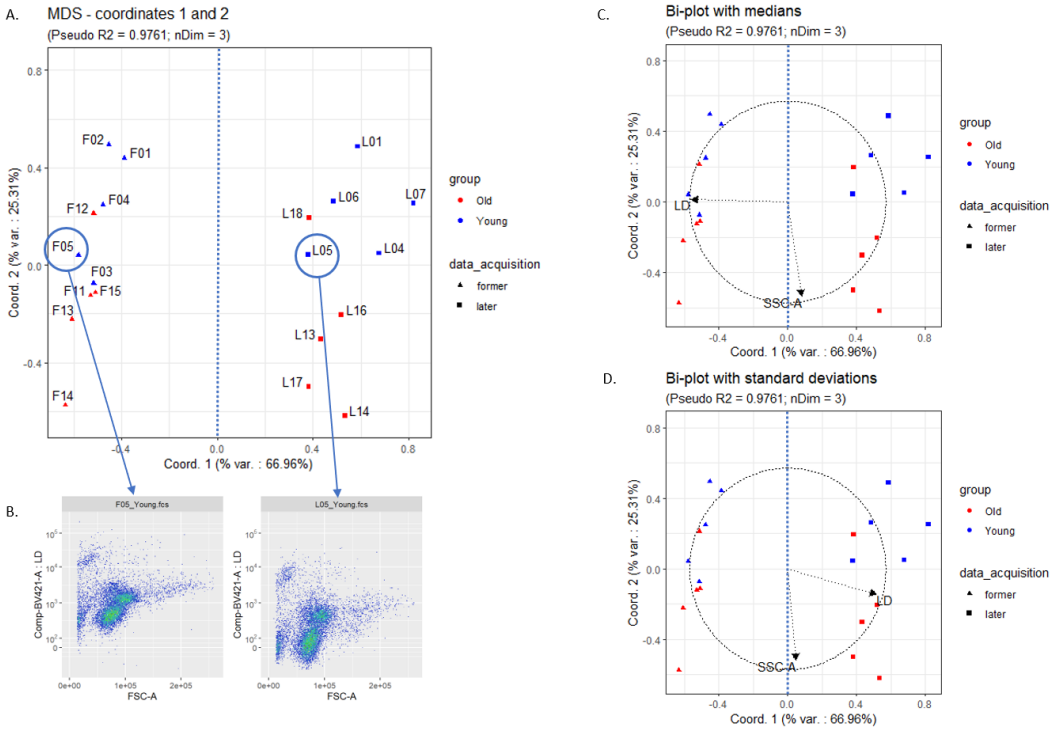

Figure S4. Low dimensional projection of the *ImmunoSenescence Human PBMC* dataset, 2D projection on axes 1 and 2. The projection plot (A) reveals a clear separation, along the first axis, between the samples of the former data acquisition (triangles), and the samples of the later data acquisition (squares). This is an obvious batch effect. According to the bi-plots using channel medians (C) and channel standard deviations (D), we can attribute this batch effect mainly to location and scale of *Live/Dead* channel. This is confirmed when plotting 2D representations of selected samples having opposite first coordinates (B).

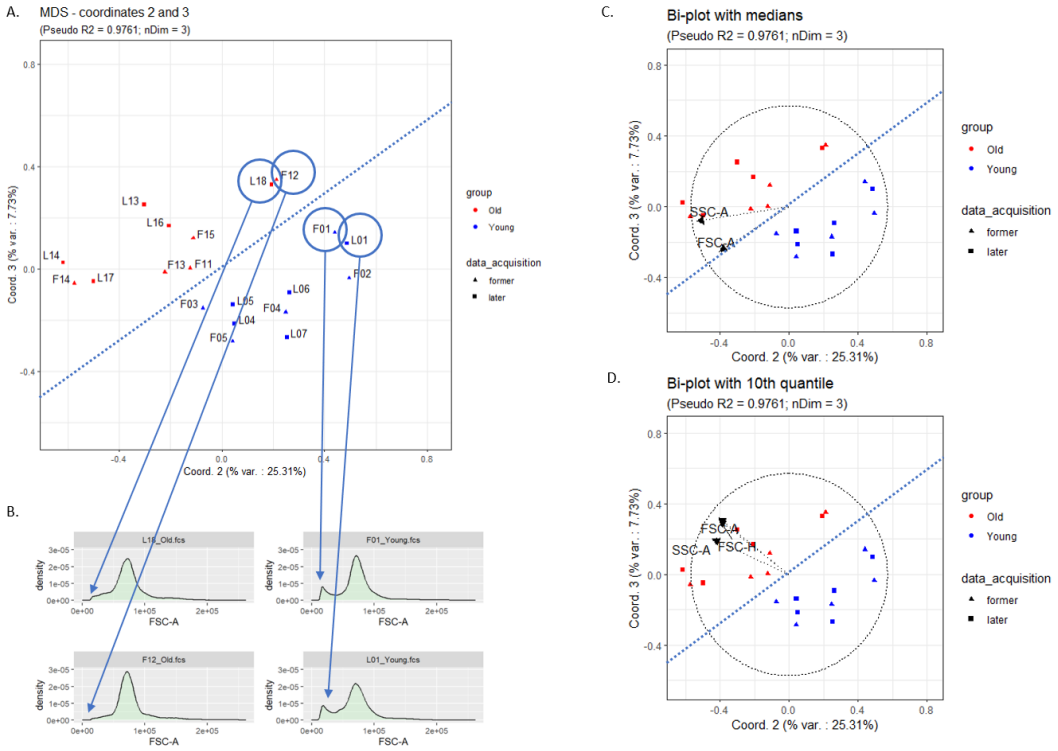

Figure S5. Low dimensional projection of the *ImmunoSenescence Human PBMC* dataset, 2D projection on axes 2 and 3. On the projection plot (A), the two groups ('old' in red vs. 'young' in blue), can be separated along a direction that is a linear combination of axes 2 and 3. Bi-plots using channel medians (C) does not highlight any channel median strongly correlated with the direction of separation. However, a bi-plot with a well-chosen statistics, i.e. the 10th percentile, enables the identification of a strong correlation with this quantile for the *FSC* and *SSC* channels. This can be confirmed when comparing the marginal distributions of the *FSC-A* channel for selected samples from the two sides of the border (B).

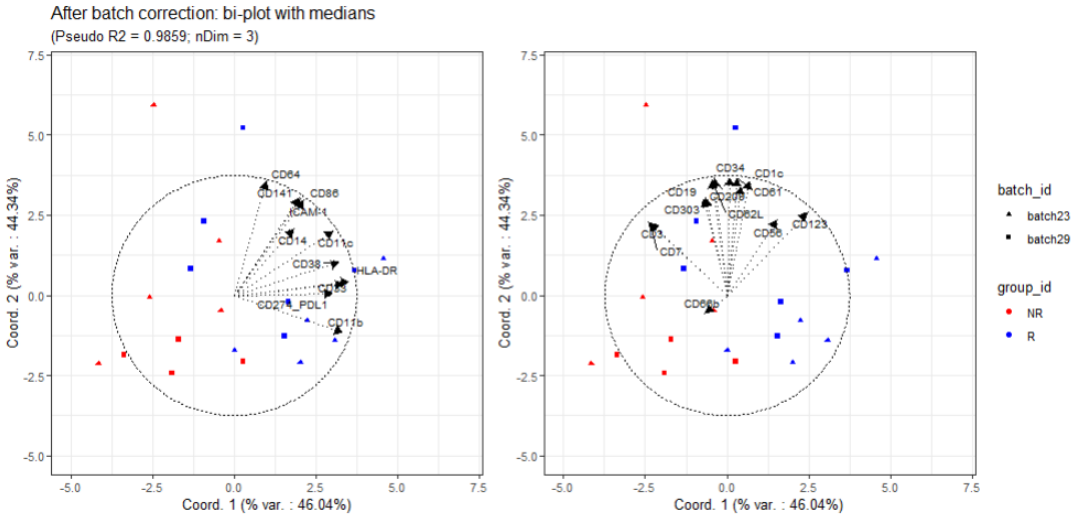

Figure S6. Bi-plots of the projection of the *Krieg_Anti_PD1* dataset on the first two axes, after batch correction, and representing all channel medians, split into two groups. In the original study (Krieg et al., 2018), it was highlighted that the frequency of a small subpopulation of CD14^{+} CD33^{+} HLA-DR^{hi} ICAM-1^{+} CD64^{+} CD141^{+} CD86^{+} CD11c^{+} CD38^{+} PD-L1^{+} CD11b^{+} monocytes in baseline samples was positively associated with responder status following immunotherapy treatment. On the left plot, we selected these 11 markers and on the right plot, the other 14 markers are represented for comparison. The intensity of the markers of the first group tends to be more correlated with the direction separating responders (blue) and non responders (red).

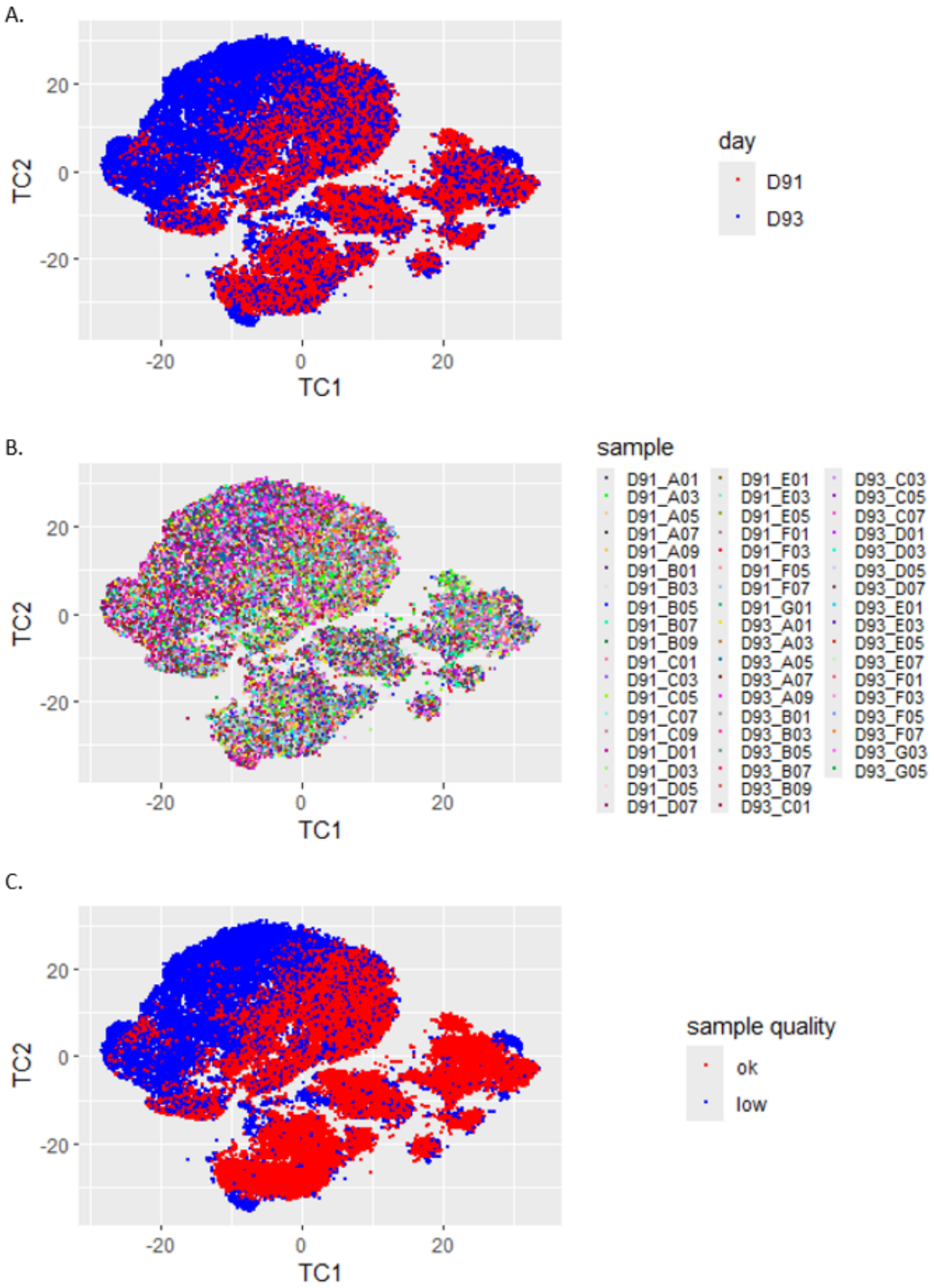

Figure S7. tSNE of the *HBV Chronic Mouse* dataset (sampling of 30,000 random events from the 55 samples). Highlighting the displayed events by acquisition day (A) suggests the existence of a corresponding batch effect. However a substantial amount of the blue dots are mixed with the red ones, because the effect is linked to a quality issue affecting only a subset of D93 samples. Highlighting the displayed events by sample (B) is uninformative, due to the high number of samples, and high number of events per sample. (C) Colouring the events by quality cluster - which was identified thanks to *CytoMDS* analysis - seems to show less residual dots mixing than in (A).

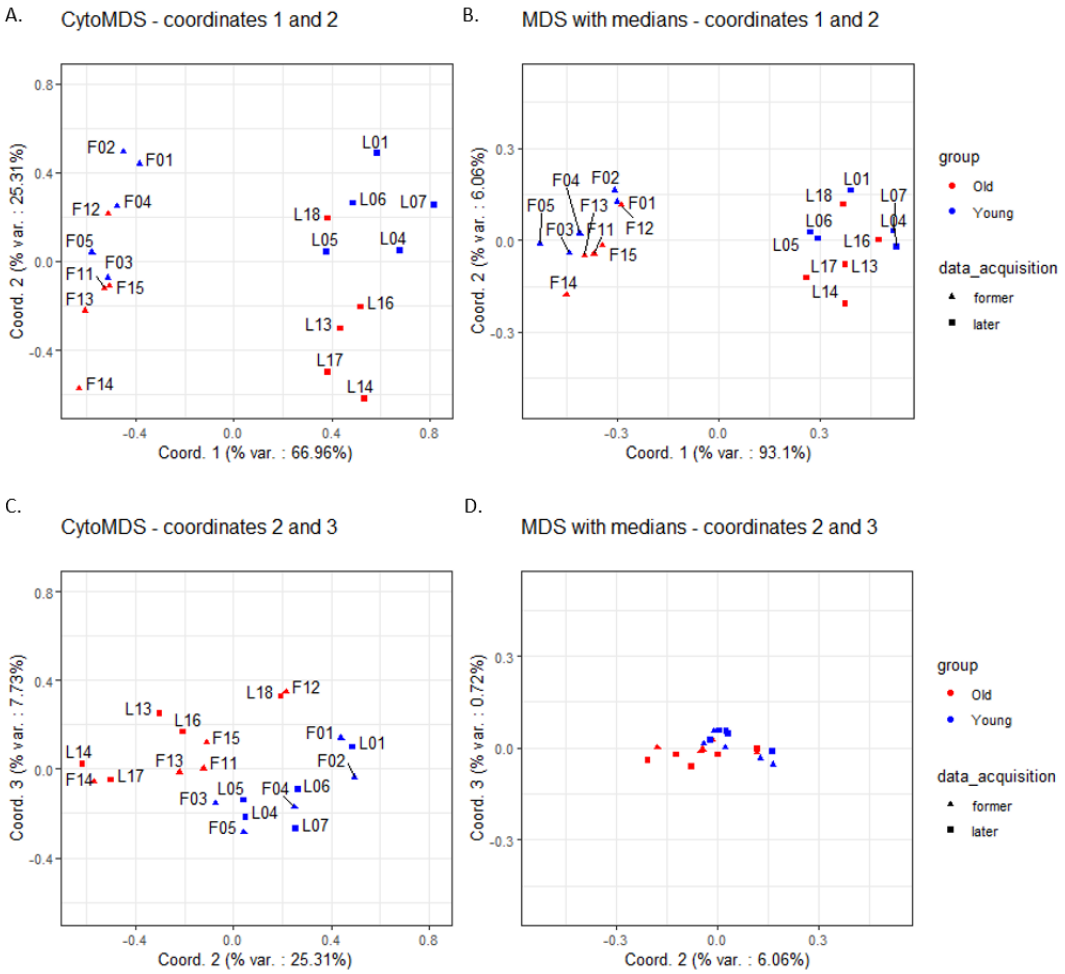

Figure S8. Low dimensional projections of the *ImmunoSenescence Human PBMC* dataset. The above plots show a comparison of the projection on axes 1 and 2 of respectively the *CytoMDS* method (A) - which uses *Earth Mover's Distances* to assess dissimilarity between samples - and *MDS* using channel medians as input (B). The two plots show similar patterns. Similarly, plots (C) and (D) show the same comparison, now using axes 2 and 3. Here, *CytoMDS* projection (C) shows a visible separation between the two sample groups. However, on the right plot (D), there is little variability left between data points. This difference can be explained by the fact that the direction of separation between the two sample groups in the *CytoMDS* projection could be associated to the 10th quantile of the *FSC* and *SSC* channel distributions (Figure S5), but not to channel medians.