BD FACSymphony A5 SE
Special Order Research Product

Received: 29 September 2020 | Revised: 28 February 2021 | Accepted: 19 July 2021

DOI: 10.1002/cyto.a.24491

COMPUTATIONAL ARTICLE

CYTOMETRY Journal of Quantitative Cell Science PART A

# Transformation of multicolour flow cytometry data with OTflow prevents misleading multivariate analysis results and incorrect immunological conclusions

Rita Folcarelli[1] | Selma van Staveren[2,3] | Gerjen Tinnevelt[1,3] | Emily Cadot[1] | Nienke Vrisekoop[2] | Lutgarde Buydens[1] | Leo Koenderman[2] | Jeroen Jansen[1] | Oscar F. van den Brink[3]

[1]Analytical Chemistry, Institute for Molecules and Materials, Radboud University Nijmegen, Nijmegen, The Netherlands

[2]Department of Respiratory Medicine, Center for Translational Immunology, UMC Utrecht, Utrecht, The Netherlands

[3]TI-COAST, Amsterdam, The Netherlands

**Correspondence**
Rita Folcarelli, Analytical Chemistry, Institute for Molecules and Materials, Radboud University, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands.
Email: r.folcarelli@science.ru.nl

## Abstract

The rapid evolution of the flow cytometry field, currently allowing the measurement of 30–50 parameters per cell, has led to a marked increase in deep multivariate information. Manual gating is insufficient to extract all this information. Therefore, multivariate analysis (MVA) methods have been developed to extract information and efficiently analyze the high-density multicolour flow cytometry (MFC) data. To aid interpretation, MFC data are often logarithmically transformed before MVA. We studied the consequences of different transformations of flow cytometry data in datasets containing negative intensities caused by background subtractions and spreading error, as logarithmic transformation of negative data is impossible. Transformations such as logicle or hyperbolic arcsine transformations allow linearity around zero, whereas higher (positive and negative) intensities are logarithmically transformed. To define the linear range, a parameter (or cofactor) must be chosen. We show how the chosen transformation parameter has great impact on the MVA results. In some cases, peak splitting is observed, producing two distributions around zero in an actual homogeneous population. This may be misinterpreted as the presence of multiple cell populations. Moreover, when performing arbitrary transformation before MVA analysis, biologically relevant and statistically significant information might be missed. We present a new algorithm, Optimal Transformation for flow cytometry data (OTflow), which uses various statistical methods to optimally choose the parameter of the transformation and prevent artifacts such as peak splitting. Arbitrary or unconsidered transformation can lead to wrong conclusions for the MVA cluster methods, dimensionality reduction methods, and classification methods. We recommend transformation of flow cytometry data by using OTflow-defined parameters estimated per channel, in order to prevent peak splitting and other artifacts in the data.

Rita Folcarelli and Selma van Staveren contributed equally to this work. Jeroen Jansen and Oscar F. van den Brink contributed equally to this work.

**KEYWORDS**
arcsinh | cofactor, flow cytometry, flowVS, inverse hyperbolic sine, logicle transformation, multivariate analysis, OTflow, preprocessing, transformation

## 1 | INTRODUCTION

The flow cytometry field has evolved rapidly. Nowadays commonly used flow cytometers measure up to 17 parameters per cell [1] and the most recently developed flow cytometers can potentially measure 30 to 50 parameters [2]. With this increase in measured variables per single cell, the information in the resulting datasets also increases tremendously.

Analysis of flow cytometry datasets by manual gating has essential drawbacks due to the lack of reproducibility, the subjectivity and bias, the inefficiency for large panels and the fact that it is time-consuming for large datasets [3]. To overcome these drawbacks various multivariate analysis (MVA) methods have been developed for automated analysis of multicolour flow cytometry (MFC) data in the past decade. MVA methods can be classified into methods which use dimensionality reduction like the broadly used method viSNE [4]. Next to viSNE, other types of dimensionality reduction methods exist, which were developed to perform specific tasks in analyzing and quantifying flow cytometry data like FLOOD [5], DAMACY (for the classification of ‘control samples’ or ‘patient samples’ based on discriminative expression patterns between these groups) [6], and ECLIPSE (for automated gating of [rare] disease-related cell populations) [7]. Clustering methods like SPADE [8], FlowSOM [9] and Citrus [10] belong to a second type of MVA methods and are frequently used by researchers to identify and characterize new (disease-related) cell subsets in a fast and unbiased way. Before one can use these methods, the MFC data should be adequately preprocessed. There are many preprocessing options, and the preprocessing steps can have considerable effects on the results of the analysis. For an overview of optional preprocessing steps, we refer to the review of Saeys et al. [3].

Up until recently, logarithmically transformed axes were used when analyzing marker expressions from MFC data using conventional gating analysis software. Log transformations perform a nonlinear conversion of the output of the analog-to-digital converter. It is useful to correct for heteroscedasticity and to change skewed distributions into more symmetric, Gaussian distributed peaks [11]. In MFC, logarithmic transformation has been very useful in coping with the wide dynamic range of emissions between fluorophores. The log scale provides an informative and proper display of populations in the higher intensity range as well as of populations in the lower intensity regions. Populations with low fluorescent intensities which may be hardly discerned on a linear scale are well visible and separated on a logarithmic scale. However, negative intensities cannot be properly displayed on a logarithmic scale, since calculating the logarithm of a negative value is undefined for real numbers and only allowed for complex numbers.

Negative fluorescence intensities are physically and biologically meaningless. Nevertheless, negative values in MFC data can occur in the process of data acquisition due to adjustment of original measurements by the instrument hardware and software. Negative values might emerge from background subtractions performed by the hardware of some flow cytometers (e.g. Becton Dickinson). PMT signals can contain high levels of background signal from fluorescent light from unbound fluorophores, PMT dark current and ambient light [12]. The background signal is measured in between two events and this signal intensity is subtracted from the next event measured. In the case of a negative or dim cell, the measured fluorescence can be smaller than the background signal, due to the measurement errors. This then leads to negative values. In BD FACS Diva software, the user of the flow cytometer cannot influence the background subtraction. Additionally, compensation of spectral overlap among fluorochromes can lead to fluorescent intensities below zero [12, 13]. Another source of negative values is caused by data spreading. At a somewhat deeper level, statistical variation originates from measurement errors such as photon counting errors and binning errors [14]. Photon counting errors are associated with the quantum mechanical nature of the emitted light. The photons that are emitted by the fluorophore, will arrive at the detector at different time points, in a Poisson-distributed manner, as it is a stochastic process. This causes the signal to be heteroscedastic, having higher variance when fluorescence intensity is higher, and leads to a nonlinear spreading of properly compensated MFC data. Attempts to represent data with negative values on a logarithmic scale lead to an apparent mean that is too high, and negative data points that are squeezed onto the axes [15]. Therefore, alternative transformation methods have been developed. The logicle [15] and hyperbolic arcsine (arcsinh) [16] transformations are the most commonly used in modern flow cytometry. Both transformation methods use a combination of linear transformation for values close to zero and a logarithmic scale for larger (negative and positive) values. The transition of the linear to the logarithmic part is smoothed out. Importantly, next to visualization of MFC data in biplots, logarithmic or bi-exponential transformations are performed as a preprocessing step for the previously mentioned MVA methods. In the viSNE and SPADE algorithms hyperbolic arcsine transformation is used with a standard parameter for the entire dataset [17], despite various reports which have shown that the choice of the parameter for transformation should be dependent on the dataset, and thus may vary per fluorescence channel [16, 18].

Here we present an algorithm for *Optimal Transformation* of *flow* cytometry data, called OTflow. OTflow is an automated and validated algorithm for optimization of transformation parameters for both visualization and MVA of flow cytometry data. The algorithm combines properties of normality of the signal and stabilization of the variances among the peaks to best represent MFC data on a bi-exponential scale. Improving the flowVS method by Azad et al [18] we combined variance stabilization [19] together with Bartlett's test [20] to define the optimal parameter for transformation based on the flow cytometry dataset itself. Variance stabilization dissociates the existing correlation between the mean intensity of a cell population and the variance, which is typical of MFC measurements. Bartlett's test [20] is used as statistics to select the arcsinh transformation cofactors which leads to homogenous variances per fluorescent channel measured. As a result, cell populations with different intensities are better discerned and the corresponding peaks resemble more normal distributions. This enables comparison of phenotypical similar cell populations across multiple samples [21]. However, variance stabilization by flowVS can be performed only when two or more peaks are present per channel. In the case of a single peak per channel, the flowVS algorithm may not be suitable to find optimal cofactor transformation. In the novel algorithm OTflow, in addition to Bartlett's statistics for multiple peaks, we integrated the possibility to estimate optimal parameters for transformations also when a single peak per channel is present. In this case, Jarque-Bera statistics test [22] for normality of the peak is applied. Moreover, OTflow specifically prevents peak-splitting in its parameter optimization process. This is very important in flow cytometry data with negative values to avoid misleading interpretation of the transformed populations. As far as our knowledge goes, no other transformation method for MFC data takes this last feature into account.

In this paper we show the versatility of OTflow by applying the algorithm to complementary datasets and performing PCA [23], viSNE [4], flowSOM [9] and Citrus [10] analyses. Additionally, we compare these MVA results to the results from the same datasets transformed using hyperbolic arcsine with the default cofactor 150, or using the recently published flowVS algorithm [18]. Also, in Appendix S1, we compare the MVA results after logicle transformation with the optimized W parameters from OTflow to those obtained after logicle transformation with a calculated W value as proposed by Parks et al. [15] We demonstrate that suboptimal transformations by a poorly chosen or calculated parameter can lead to misleading MVA results that may subsequently invoke incorrect immunological conclusions.

## 2 | METHODS

### 2.1 | Transformations

We considered two of the most widespread transformations used to process flow cytometry data: inverse hyperbolic sine function [16] and the logicle function [15]. They both belong to the class of bi-exponential functions and have the characteristics of being linear and symmetric near zero and becoming exponential for higher values, with a smooth transition between the linear and exponential regions [24]. Note that both transformations deal with negative numbers. For detailed description of both functions we refer to the Supplementary Material I.

### 2.2 | Variance stabilization and normality of flow cytometry peaks

Bi-exponential function-based transformations perform variance stabilization of the signals, as they remove the correlation between data variability and mean, which is typically present in MFC data. If the variance between the signals is not stabilized, cell populations with higher signal intensity will have a larger variance which does not necessarily reflect the true marker variability. For accuracy in MVA, optimally transformed MFC data is characterized by peaks resembling normal distributions with homogenous variances (homoscedasticity), which are not dependent on the fluorescence intensity. The variance between distributions, after optimal transformation, will therefore reflect the true variability in protein expression of the cell subsets. This facilitates the comparison of cell populations with different marker expression levels. Checking for variance stabilization between multiple peaks for one marker can be used as measure of how well (bi-exponential) functions transform the MFC data. This may be evaluated by using Bartlett's likelihood-ratio test, commonly chosen in statistics to check for homoscedasticity among multiple groups. The Bartlett's test is adopted in the flowVS algorithm [18, 20]. Also in OTflow, we use Bartlett's test to assess the homogeneity of variance between peaks per measured marker across all the individuals. However, when a single cell population is present per marker, variance stabilization cannot be applied to peaks within the same sample, but only to peaks across all the individuals. With a single peak per sample and a low number of samples present, optimizing Bartlett's statistics is not preferred, since biologically relevant variance between different samples might be removed. Alternatively, optimal transformation of the single peak can be evaluated by how well the transformed peak resembles a normal distribution. The Jarque-Bera test is used to determine the normality of the single peaks. The test is based on estimation of kurtosis and skewness of the peak, which are schematically represented in Figure S2. These are properties related to the shape of distributions and they are commonly used to check the deviation from normality.

### 2.3 | The algorithm step by step

A schematic overview of the algorithm is shown in Figure 1.

We describe below the process to choose an optimal cofactor for arcsinh transformation (Equation S1) for each channel measured. The same process can be applied to the optimization problem of the parameter W for logicle transformation (Equation S2).

**Step 0**: *Multiset structure of MFC data*. The OTflow algorithm is simultaneously applied to all the MFC samples present in the study. Single MFC samples can be arranged in the ‘multiset’ matrix X, of size $\left(\sum_{1}^{I} N_i \times J\right)$, where $N_i$ is the number of cells of the $i^{\text{th}}$-individual and $J$ corresponds to the markers measured, 1…$j$…$J$.

**Step 1**: *Data transformation*. The matrix X is transformed by the arcsinh function as described in Equation S1. We define $\mathrm{X}_{\log_i} = \mathit{arsinh}\left(\frac{\mathrm{X_i}}{c}\right)$ the arcsinh transformed matrix of the i-th sample, of size $(N_i \times J)$. The value of cofactor $c$ is progressively increased by assuming values $e^a$, with $a$ ranging from 0 to 10, in steps of 0.05, corresponding to 201 cofactor values.

**Step 2**: *Each marker is analyzed per individual*. Each sample $\mathrm{X}_{\log_i}$ is randomly subsampled (without replacement) to 1000 cells. The subsampling is repeated 100 times by a Monte-Carlo cross-validation. If the sample size is smaller than 1000 cells, then the nonsubsampled set is used. For each subsample, the algorithm is applied to each marker individually, included in the column vector $\mathbf{x}_{\log_{i,j}}$, representing the arcsinh-transformed $j^{\text{th}}$-channel of the $i^{\text{th}}$-individual.

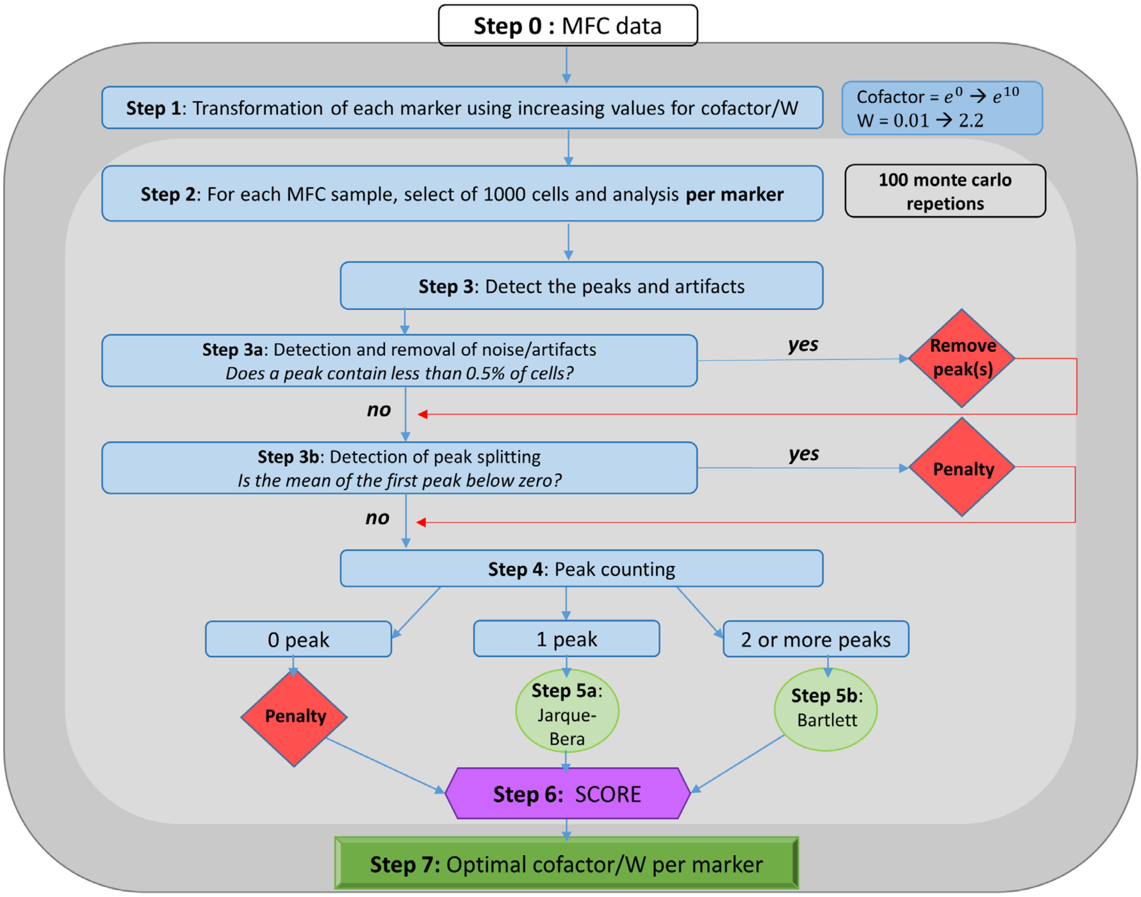
Step 0 : MFC data
Step 1: Transformation of each marker using increasing values for cofactor/W
Cofactor = $e^0$ → $e^{10}$
W = 0.01 → 2.2
Step 2: For each MFC sample, select of 1000 cells and analysis per marker
100 monte carlo repetions
Step 3: Detect the peaks and artifacts
Step 3a: Detection and removal of noise/artifacts
Does a peak contain less than 0.5% of cells?
yes
Remove peak(s)
no
Step 3b: Detection of peak splitting
Is the mean of the first peak below zero?
yes
Penalty
no
Step 4: Peak counting
0 peak
1 peak
2 or more peaks
Penalty
Step 5a: Jarque-Bera
Step 5b: Bartlett
Step 6: SCORE
Step 7: Optimal cofactor/W per marker

FIGURE 1 Schematic step-by-step representation of the OTflow algorithm. Steps 2–6, included in the light gray sector, are repeated through the 100 monte-carlo iterations. The validated final score is calculated as average of the score obtained each repetition. The score is a result of Jarque-Bera statistics (estimated for single peak), Bartlett statistics (estimated for 2 or more peaks) and Penalty. A penalty is given when a negative peak is found (peak splitting-Step 3b) or no peak is detected (Step 4). The optimal cofactor or W parameter (Step 7) is the one that produces the lowest mean score per marker

**Step 3**: *Peak detection*. Peaks or cell subpopulations are determined for each separate marker per measurement. First univariate probability density is estimated with kernel density estimate (KDE) [25] by using Gaussian function as kernel. The probability density function (PDF) is used to describe the probability that a cellular marker expression assumes a certain value, which can be established by KDE. The estimate basically fits a smooth curve (specifically a Gaussian) on the data as a kind of continuous replacement for the discrete histogram.

Then a peak finding algorithm (*findpeak* function in Matlab [26]) is applied to the resulting density estimate. Peaks are identified by using a minimal peak prominence of 0.1% of the density estimate. The prominence of a peak measures how much the peak emerges due to its absolute height (e.g. number of cells) and its location or distance relative to other peaks. If this is bigger than the chosen threshold for the prominence, separate peaks will be identified.

**Step 3a**: *Detection and removal of noise/artifacts*. Peaks containing less than 5 cells will be disregarded from further analysis. This will prevent inclusion of very small peaks in the analysis which might be associated with outliers or noise in the data.

**Step 3b**: *Detection of peak splitting*. The mean of each leftmost peak is estimated. In case a negative mean is detected, a penalty is given to the measurement, as this suggests the presence of *peak splitting* due to the transformation.

**Step 4:** *Peak counting and penalties*. Subsequently, the number of peaks is checked per measurement for each marker. Depending on the number of peaks found, different statistics or penalties are applied. **a)** If no peaks are present (which means all the detected peaks contained less than 5 cells), a penalty is given; **b)** if only one peak is present, *Jarque-Bera* statistics are applied; **c)** Samples containing multiple (two or more) peaks are grouped together; *Bartlett* statistics are then applied to check variance stability of all the peaks from the merged samples.

**Step 5a:** *Jarque-Bera statistics*. When a single peak per marker is found, the Jarque-Bera (*JB*) statistical method [27] is applied to optimize the normality of the peak. For each value of the cofactor *c*, the score JB(*c*) is estimated as shown in Equation 1.

$$\begin{aligned}
&(a)\ \mathrm{S}(c)=\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{3}{2}}}\right)\\
&(b)\ \mathrm{K}(c)=\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{2}}\right)\\
&(c)\ \mathrm{JB}(c)=\frac{n}{6}\left(S^2+\frac{(K-3)^2}{4}\right)
\end{aligned}\tag{1}$$

Where $\mathrm{S}(c)$ and $\mathrm{K}(c)$ represent skewness and kurtosis of the distribution of the peak, respectively. The smaller the value of the JB(c) statistics, the more likely the peak resembles a normal distribution.

**Step 5b:** *Bartlett statistics*. Subsamples with multiple peaks per marker are bundled together. Bartlett statistics test for equal variance across all the peaks from all the merged subsamples against the alternative hypothesis that variances are unequal across the peaks. The Bartlett statistics are calculated for each of the 201 values of $c$ (see Equation 2), as follow:

$$\text{Bartlett}\,(c) = \frac{(N-k)\ln\left(\sigma_p{}^2\right) - \sum_{i=1}^{k}(n_1-1)\ln\left(\sigma_i{}^2\right)}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right)-\frac{1}{N-k}\right)} \tag{2}$$

where $N$ is the number of cells included in all the found peaks, $k$ is the number of peaks, $\sigma_p{}^2$ the pooled variance of peaks, $\sigma_i{}^2$ is the variance of the $i^{\text{th}}$-peak. When the variances between the peaks are equal, the Bartlett (c) assumes the lowest value.

**Step 6–7:** *Final score and optimal cofactor*. For each Monte-Carlo iteration a score is calculated as contribution of penalties, *Jarque-Bera* and *Bartlett* statistics. The two statistical tests are weighted to assure that both statistics have similar impact in determining the final score. Mean and standard deviation of the scores of the 100 Monte-Carlo iterations are estimated. The optimum cofactor is the value corresponding to the lowest mean scores for each marker.

## 2.4 | OTflow code online

The algorithm is shared online: https://surfdrive.surf.nl/files/index.php/s/zTzLY1YlPzKuNWo.

## 2.5 | Data sets used to test OTflow

The parameter optimization algorithm OTflow was applied to a HIV dataset consisting of seven healthy control samples and 18 patient samples, of the patients 14 were treated with anti-retroviral therapy and four were untreated. Peripheral blood samples were obtained from individuals in the context of standard diagnostic care. Residual blood not used for the standard diagnosic tests was used for research, with Informed Consent from patient according to protocols of the UMC Utrecht.

The MFC panel comprised 10 fluorescently labeled antibodies: CD123, CD14, CD16, CD3, CD4, CD8, CD56, CD20, CD193, CD62L for identification of different leukocyte subsets in the peripheral blood. After measuring single stains per antibody and a negative control using white blood cells and Compensation beads (BD CompBead Anti-Mouse Ig,κ/ Negative Control Particles Set) a compensation matrix was generated using Flowjo analysis software (Tree Star Inc., Ashland, Oregon). Per sample between 250,000–500,000 cells were measured on a BD LSR-II Fortessa. Prior to MVA, the single cells were selected, and debris was excluded from the analysis (Figure S7) using manual gating (Flowjo analysis software, Tree Star Inc., Ashland, Oregon). The gated fcs files were exported.

The OTflow algorithm was tested on two additional flow cytometry datasets with varying number of cell subpopulations per dataset, as described in the Supplementary Material I.

## 2.6 | Simulation study to test OTflow

A simulation study was performed with univariate data to visualize and quantify the behavior of the OTflow algorithm, and compare it with flowVS. The simulation study is further described in the Supplementary Material V.

## 2.7 | MVA

Here the way different multivariate analysis methods were applied is shortly explained. For background information about the methods we refer to the Supplementary Material II or the original papers about the methods (see references).

### 2.7.1 | Principal component analysis

Principal component analysis [23] (PCA) is a dimensionality reduction technique which has been widely used for analysis of MFC data [6, 7, 28, 29]. After arcsinh transformation, the data were mean centered and scaled over all the samples, prior to the PCA analyses.

### 2.7.2 | viSNE

viSNE analyses were performed in Cytobank [17] with the following parameters: perplexity 30, random seed, # Iterations 1000, Theta 0.5. To perform the viSNE analysis on the HIV dataset, each individual was randomly down-sampled to 2000 events. This enabled decrease of the running time and thereby computational burden and crowding effects were avoided.

### 2.7.3 | FlowSOM

The flowSOM [9] algorithm was performed by running the R code available at Bioconductor. The algorithm was trained on the MFC data using suggested (default) parameters: 100 as number of nodes, Euclidean distance to find nearest neighbor, and a training length of 10 epochs. A minimum spanning tree was then built to visualize the clusters detected by the algorithm. The FlowSOM algorithm was run by using a fixed random seed (=25), this will allow to get the same exact flowSOM minimum spanning tree when repeating the analysis on the same data. Thereby, differences found in flowSOM results between the differently transformed datasets are with certainty caused by the transformations, and are not a consequence of the stochastic nature of the method.

#### 2.7.4 | Citrus

Citrus [10] was run using the online platform Cytobank [17] with the following default parameters: minimum cluster size of 5% and a cross-validation fold of five.

## 3 | RESULTS

### 3.1 | Flow cytometry datasets contain a considerable amount of negative values, requiring careful attention when transforming the data

The datasets used in this article have a percentage of cells displaying one or more negative intensities ranging between 39% and 96%; these negative intensities vary by eight or nine orders of magnitude (Table S1).

As explained in the introduction, subtraction of the background signal may be one of the causes of negative values in a dataset. If a lot of unbound fluorophores are present in the solution, the background signal can be higher than the signal of a particle or cell to which no such antibody has bound, again leading to negative values (Figure S3).

Another cause of the appearance of negative intensities in a flow cytometry dataset, is data spreading. Cell populations that are negative for a certain marker (low fluorescence intensity), but positive for another marker (high fluorescence intensity) with spectral overlap in the channel of the negative marker, can cause data spreading for the negative marker. When compensation is applied, the intensities for the negative marker will be decreased and part of the lowest values are shifted into the negative range (Figure S4) [30].

For flow cytometry data, the log scale provides an informative and proper display of populations in the higher intensity range as well as of populations in the lower intensity regions. Populations with positive low fluorescent intensities which may be hardly discerned on a linear scale are well visible and separated on a logarithmic scale (-Figure S5). However, negative values cannot be logarithmically transformed. In Figure S6 differently transformed scales are shown for the visualization of flow cytometry data. When data which contains negative intensities are represented on a logarithmic scale, the originally negative values are plotted onto the axes or lost in the representation (Figure S6B). To overcome this problem, it might be suggested to shift the whole dataset to positive values and then apply log transformation. Doing so, events in the lower region are overly dispersed and the ratio between values changes. This would lead to incorrect estimation of signal means (Figure S6C). As a solution bi-exponential transformations are used, e.g. arcsinh or logicle transformation. For both functions, the width of the linear region is determined by a parameter. When this parameter is not carefully chosen, artifacts can arise at the transition point of the linear part to the logarithmic part [31]. Figure S6D–G shows logicle and arcsinh transformed flow cytometry data with different parameters. Homogeneous populations with a mean intensity close to zero can be split into two distinct populations when the value for the parameter is too low. We refer to this phenomenon as peak splitting. In Figure S6D,F suboptimal parameters were used for arcsinh and logicle transformation, respectively. In these plots 6 and 4 or 5 cell populations can be discerned, respectively, while in the figures with an optimally defined parameter for transformation (Figure S6E,G) 3 cell populations are found. The newly developed algorithm, named OTflow, enables to define the optimal co-factor or W value for arcsinh or logicle transformation, respectively, with the aim to prevent artifacts such as peak splitting. The algorithm has been described in the methods section step by step.

### 3.2 | Transformation applied to a real dataset: HIV Patient data

The HIV patient and control dataset comprises of blood samples obtained from seven healthy controls and 18 HIV patients. The patient group is heterogeneous, as it includes both treated and untreated patients. The MFC panel used contained 10 surface markers, mostly differentiation markers, aimed to identify the most common leukocyte subsets in the peripheral blood. Doublets and cellular debris were excluded from the dataset by manual gating (-Figure S7). OTflow was used to identify the optimal cofactor for the arcsinh transformation for each of the 10 fluorescence channels in the HIV patient and control dataset. Bartlett or Jarque-Bera statistics were estimated by the OTflow algorithm depending on the number of peaks found per cofactor tested, ranging between $e^0$ and $e^{10}$ (=1 and ~22,026). Additionally, a penalty was given when no peaks were present or when negative peaks were found in the transformed data. The average final score and standard deviation obtained per marker for varying cofactors are shown in Figure S8. For most of the markers, higher scores are obtained when very low and very high values of the cofactor are used for the arcsinh transformation, while for the medium values a dip in the curves is present between $e^4$ and $e^7$ (≈54.6–1097). An exception to this trend is observed for marker CD20 having maximum score value for central cofactors. By investigating the contribution of the different statistics (Bartlett's and Jarque-Bera) to the average score (Figure S9), we observed that such trend is due mainly to Bartlett's statistics, meaning that variances between peaks in this range of cofactors are not well stabilized.

Table 1 compares the optimal cofactors found by OTflow with the results obtained by applying flowVS on the same dataset.

Running the flowVS algorithm on the dataset led to significantly lower cofactors for marker CD123, CD16, CD56 and CD20 compared to the results of OTflow and to the default cofactor 150. Figure2 A-C shows the distribution of these four representative surface markers on leukocytes of healthy controls (blue) and HIV patients (red), after transformation with a default cofactor 150 (Figure 2A), or with the cofactors determined by flowVS (Figure 2B) or OTflow (Figure 2C),

**TABLE 1** Cofactors for arcsinh transformation calculated per marker by flowVS and OTflow algorithms

| | CD123 | CD14 | CD8 | CD4 | CD3 | CD16 | CD62L | CD193 | CD56 | CD20 |
|---|---|---|---|---|---|---|---|---|---|---|
| flowVS | 20 | 1233 | 150 | 1164 | 186 | 12 | 359 | 132 | 0.3 | 90 |
| OTflow | 116 | 944 | 314 | 665 | 245 | 735 | 632 | 172 | 572 | 734 |

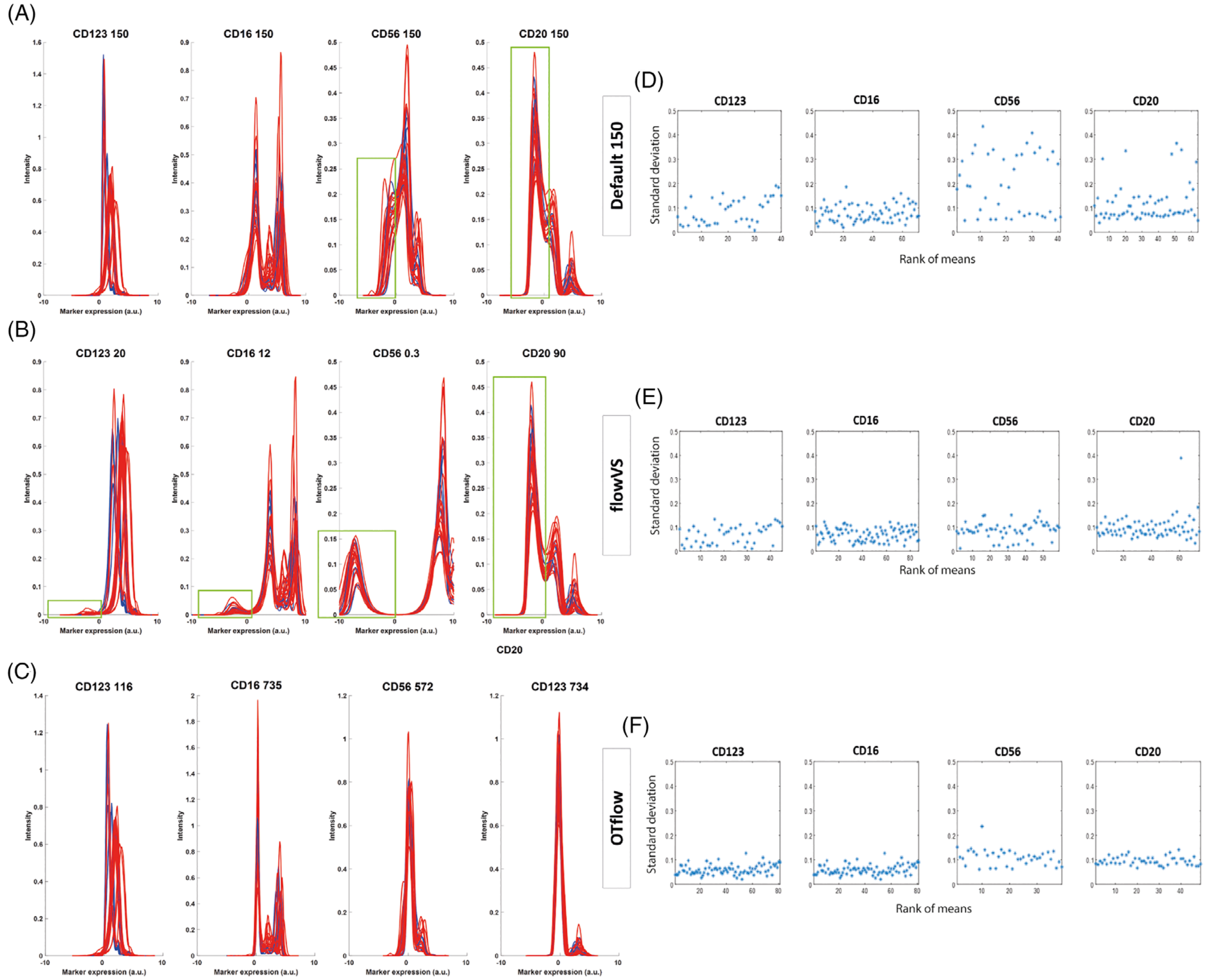

**FIGURE 2** In contrast to OTflow based transformation, arcsinh transformation with default cofactor 150 and flowVS based transformation leads to peak splitting and OTflow performs better regarding stabilization of the fluorescence signal. (A–C) Histograms of the leukocyte expression levels of CD123, CD16, CD20 and CD56 per sample after applying various transformation methods – (A) transformation with default cofactor 150, (B) with flowVS defined cofactors, (C) with OTflow defined cofactors. Blue lines represent the 7 control samples, while the red lines represent the 18 patient samples. Negative peaks originated due to suboptimal transformation, resulting into peak splitting, are marked with a green box. (D–E) The standard deviation of the peaks plotted against the rank of MFI for arcsinh transformation with (D) default cofactor 150, (E) with flowVS defined cofactors, (F) with OTflow defined cofactors

respectively. Arcsinh transformation with the default cofactor of 150 for all markers, produces spurious splits of the cell population around 0 for CD56 and CD20 (Figure 2A, marked with green rectangles), creating 2 peaks which are partly overlapping. When using the cofactors calculated by flowVS, negative intensities, resulting from peak splitting, were present for all the four selected markers (Figure 2B, marked in green). The emergence of the extra peaks in the data could lead to the conclusion that there are three populations present: a cell population not expressing the marker; a cell population with a weak expression of the marker and a cell population which strongly expresses the marker. Transformation with OTflow-defined cofactors did not lead to peak splitting (Figure 2C). For the remaining markers no peak splitting was present after transformation with default cofactor 150 and cofactors determined by flowVS, as shown in Figure S10. When using the optimal cofactors from OTflow for the transformation, the variance between the peaks is stabilized more evenly for all the measured markers, especially compared to the transformation with the default cofactor.

Stability of variance performed by the transformation methods used was quantitatively evaluated and the results are shown in Figure 2D–F. In a similar way as done in [18], we calculated the standard deviation of the peaks identified for marker CD123, CD16, CD56 and CD20, after transformation and range normalization. The standard deviations per marker were then plotted against the rank of the mean fluorescence intensity of the peak. The rank of the means instead of actual means was used, to distribute the points evenly along the x-axis. For most of the markers, all the three transformation methods show a small and stable standard deviation for most of the peaks, meaning that transformations are able to correct for peak heteroscedasticity (standard deviation is not dependent on the fluorescence intensity of the peak). Exceptions are present for marker CD56 and CD20 when arcsinh transformation is applied using standard cofactor 150, which shows large differences in the standard deviation for the different peaks. flowVS and OTflow methods are overall comparable in stabilizing the variance of the peaks, with OTflow performing slightly better in stabilizing the variance of marker CD20. Importantly to mention, as shown already in Figure 2B, in contrast to OTflow transformation, flowVS-based transformation leads to peak-splitting and more peaks are thus identified.

The runtimes for both flowVS and OTflow transformation on the HIV dataset were compared (Table 2). The fact that the OTflow transformation is more time consuming than the flowVS algorithm, is mostly due to the validation step of the OTflow transformation (per sample and per fluorescence channel 20 Monte Carlo iterations are performed).

OTflow was applied also to find the optimal *W* parameter for the logicle transformation. The OTflow-based transformation results can be found in the Supplementary Material III. These were compared to the results obtained by using the calculated *W* with Equation S2b, which is the same function used in the estimatelogicle in the flowCore package [32, 33]. From this, we concluded that also for logicle transformation peak splitting is prevented when using the OTflow calculation, leading to more reliable results. The OTflow-based transformation results were also compared with the results obtained when using the *W* as estimated by the FCStrans method [34] (-Figure S12C). FCStrans-based transformation also shows peak splitting of the left-most peak for CD56 expression. Secondly, FCS transformation leads to over transformation of some markers, resulting in very spiky peaks (Figure 12C CD123, CD8, CD3, CD193). Finally, when comparing the three methods, FCStrans performs worse in variance stabilization of CD3 expression levels.

**TABLE 2** Runtime comparison of flowVS and OTflow algorithm for the HIV dataset

| | Runtime (s) |
|---|---|
| flowVS | 409 |
| OTflow | 6466 |

## 3.3 | Multivariate analysis

We evaluated the effect of the arcsinh transformation using the default cofactor 150 and cofactors determined by OTflow and flowVS algorithms on the HIV dataset by applying various multivariate analysis methods which enable visualization of MFC data. These methods include the dimension reduction techniques Principal Component Analysis (PCA) [23] and viSNE [4]; the clustering method flowSOM [9] and the classification method Citrus [10].

### 3.3.1 | Principal component analysis on the transformed HIV dataset

When analyzing the PCA results of the differently transformed data, differences were observed in the distribution of the loadings and in the cell score distributions (Figure 3). The order in the loadings of CD123, CD56 and CD16 differs, with the greatest differences in direction for CD123 and CD56. Focusing on the scores, in the PCA models of the flowVS-transformed data, two putatively CD3$^{bright}$CD4$^{bright}$ and two CD3$^{bright}$CD8$^{bright}$ cell subsets were identified next to each other in healthy controls by using the vectors of the markers as a compass for marker co-expression level (Figure 3B). When analyzing the PCA model of the optimally transformed data by OTflow, also 2 types of CD3$^{bright}$CD4$^{bright}$ populations and 2 types of CD3$^{bright}$CD8$^{bright}$ populations were found (Figure 3C), one major population and below the major population one minor population attached to it. The PCA results of the datasets transformed with default cofactor 150 shows no additional population of CD3$^{bright}$CD4$^{bright}$cells or CD3$^{bright}$CD8$^{bright}$ cells (Figure 3A). All these (sub-)populations were further explored through backgating which revealed a different marker co-expression of the CD3$^{bright}$CD4$^{bright}$ and CD3$^{bright}$CD8$^{bright}$ subsets, depending on the transformation method used.

The backgating results of the population of the flowVS transformed data are shown in Figure S13. The two CD3$^{bright}$CD4$^{bright}$ subsets differed only based on CD56 expression levels: one of the subsets having almost only negative intensities for CD56 while the other subset mostly consisted of positive intensities for CD56 (-Figure S13A,B). The same difference accounts for the two CD3$^{bright}$CD8$^{bright}$ subsets (Figure S13A,C). The separation of cells with positive and negative expression levels for CD56 is caused by peak splitting. A very low cofactor of 0.3 was defined by the flowVS algorithm (Table 1), leading to a wide linear range around 0 after transformation and thus resulting into peak splitting when visualizing the data in a biexponentional graph. In contrast, the CD3$^{bright}$CD4$^{bright}$ and CD3$^{bright}$CD8$^{bright}$ cells of the data transformed with cofactor 150, or with cofactor 572 as defined by OTflow (Table 1), both have low expressions for CD56 (data not shown).

The cells with positive intensities for CD56 after using the flowVS defined cofactor could be interpreted as CD3$^{bright}$CD4$^{bright}$CD56$^{bright}$ and CD3$^{bright}$CD8$^{bright}$CD56$^{bright}$ NKT-like cells based on the loading position of CD56 and on the backgating of these populations. The presence of these cell subsets in the peripheral blood of healthy subjects has

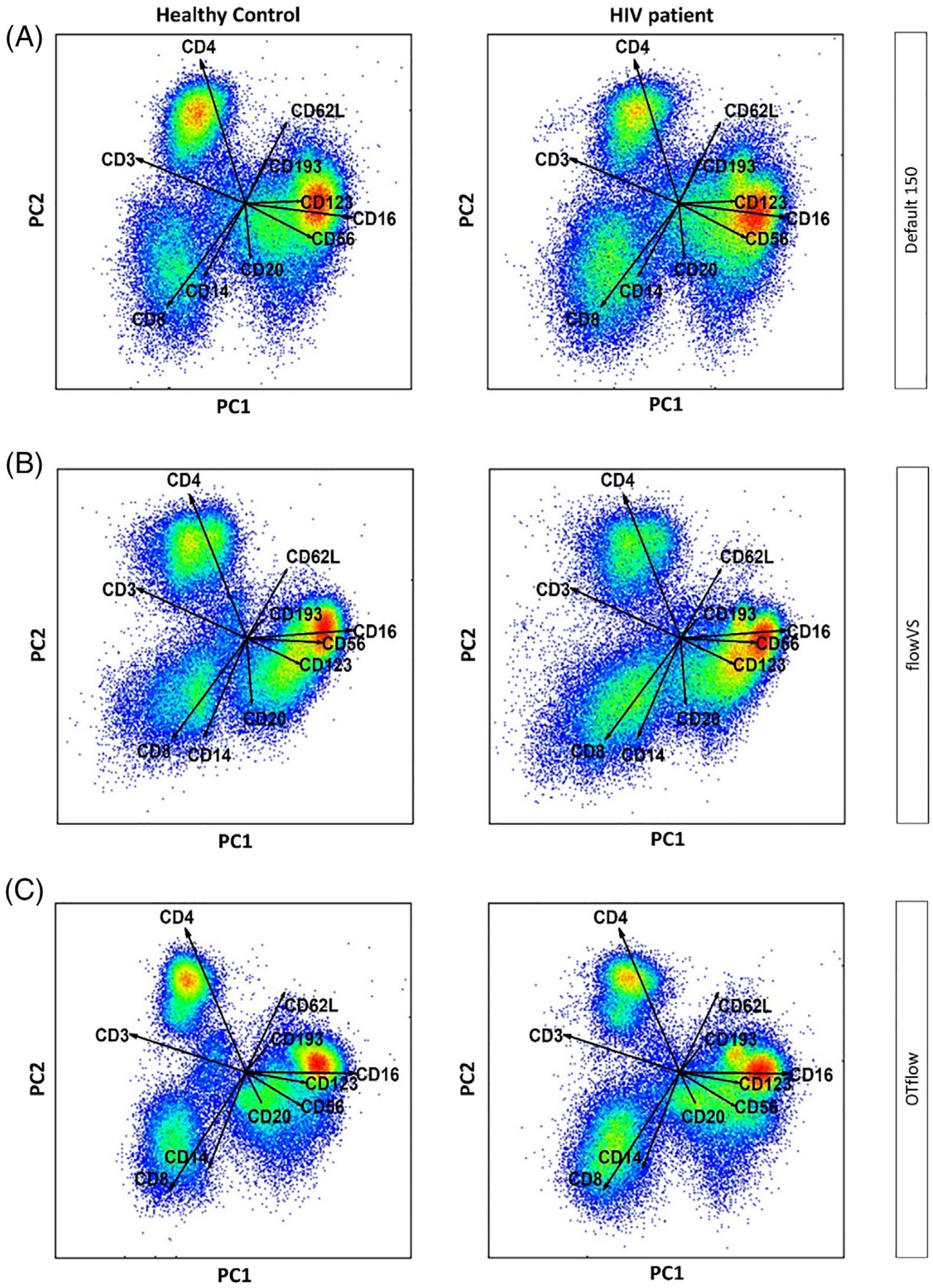

**FIGURE 3** PCA of differently transformed HIV datasets. PCA plots of PC1 vs PC2 of the HIV data after arcsinh transformation with default cofactor 150 (A), flowVS determined cofactors per channel (B), and channelspecific cofactors as calculated by OTflow (C). The cells are plotted based on the PC1 and PC2 scores and the loadings (vectors), representing the marker expressions, are projected within the PC models. The left panels represent all healthy controls and the right panels represent the HIV patients (100,000 cells are shown per plot)

been described [35], but not in these high proportions (38% of total $CD3^{bright}CD4^{bright}CD56^{bright}$ cells and 55% of total $CD3^{bright}CD8^{bright}CD56^{bright}$ cells, respectively). MVA results from these inadequately transformed data could result into misleading interpretations of the findings, when drawing conclusions about percentages of $CD3^{bright}CD4^{bright}CD56^{bright}$ and $CD3^{bright}CD8^{bright}CD56^{bright}$ populations in healthy controls vs HIV patients.

Importantly, besides the appearance of artificial cell populations, suboptimal transformation can also lead to loss of biologically relevant information. In contrast to the flowVS transformed data, the two $CD3^{bright}CD4^{bright}$ populations in the OTflow transformed analysis expressed CD56 to the same level ($CD56^{low}$, data not shown). The two $CD3^{bright}CD4^{bright}$ populations differed from each other based on expression levels of CD62L. Naïve T cells express CD62L at high levels, while memory T cells display low levels of CD62L expression. When gating these cell subsets it appeared that HIV patients had significantly higher percentages of CD3$^{bright}$CD4$^{bright}$CD62L$^{low}$ cells compared to healthy controls (Figure S14, $p = 0.018$), as has been reported before [36–38]. This difference was missed in the PCA model of the flowVS transformed data because of the more notable but artificial difference in CD56 expression for the CD3$^{bright}$CD4$^{bright}$ cells caused by peak splitting. Also in the PCA model of the transformed data with the default cofactor 150 this information was not represented, since we did not find two separate CD3$^{bright}$CD4$^{bright}$ cell populations.

### 3.3.2 | viSNE analysis on the transformed HIV dataset

The viSNE algorithm [4] was also applied to the differentially transformed HIV dataset. This showed that the deceptive results were inherent to suboptimal transformed data, and not specific for a certain type of MVA technique. The resulting viSNE maps, colored based on the expression levels of CD3, CD4, CD8, CD62L, CD14 and CD56, are shown in Figure 4. When comparing the viSNE maps of the differently transformed data, the flowVS transformed dataset contains more 'cell populations' than the datasets transformed with the default cofactor or with the cofactors defined by OTflow. Comparable to the PCA results, the viSNE map of the flowVS transformed data shows CD3$^{bright}$CD4$^{bright}$ cells and CD3$^{bright}$CD8$^{bright}$ cells divided over two clusters, of which one seems to be CD56$^{bright}$ and the other CD56$^{low}$ (Figure 4B). In addition to these putative T-cell sub-clusters, also monocytes (CD14$^{bright}$ expression) appear to be split in two separate clusters, based on high and low CD56 expression levels. Furthermore, the CD56 heat map shows a clear pattern with only CD56$^{low}$ or CD56$^{high}$ cells, while for the transformation with both the default cofactor 150 and OTflow-based cofactors, the CD56 expression levels gradually range from low to high. Both for the default 150 and OTflow transformed viSNE results the CD4$^{bright}$ and CD8$^{bright}$ T-cell populations and monocyte population are represented by a more homogeneous cell cluster, not subdivided based on CD56 diverse expression levels (Figure 4A,C–CD56 plot). In line with the PCA results, the viSNE analysis of the data transformed by flowVS also shows less distinct cell populations based on CD62L expression (Figure 4B – CD62L plot) when compared to the data transformed with default cofactor 150 or the OTflow transformed data (Figure 4A,C–CD62L plot).

viSNE is a stochastic algorithm, which means different results are generated on the same dataset when the analysis is performed twice. To show that the differences found in expression levels between the transformation methods in Figure 4 are not caused by the stochasticity of the algorithm, two additional viSNE runs were performed per transformed dataset (Figure S15). Similar expression levels of cell clusters per transformation method were found, the cell clusters are merely located at different locations in the tSNE plots. From this can be concluded that the differences observed in Figure 4 are truly caused by the differences in transformation methods.

In summary, viSNE analysis of poorly transformed data results into a more complicated viSNE map with cell populations that manifest due to peak splitting.

### 3.3.3 | flowSOM on the transformed HIV dataset

Next to PCA and viSNE analyses, which both employ dimension reduction techniques to visualize MFC data, we performed cluster analysis using the flowSOM algorithm [9]. The algorithm produces self-organizing maps to visualize MFC data in clusters/nodes represented either in a

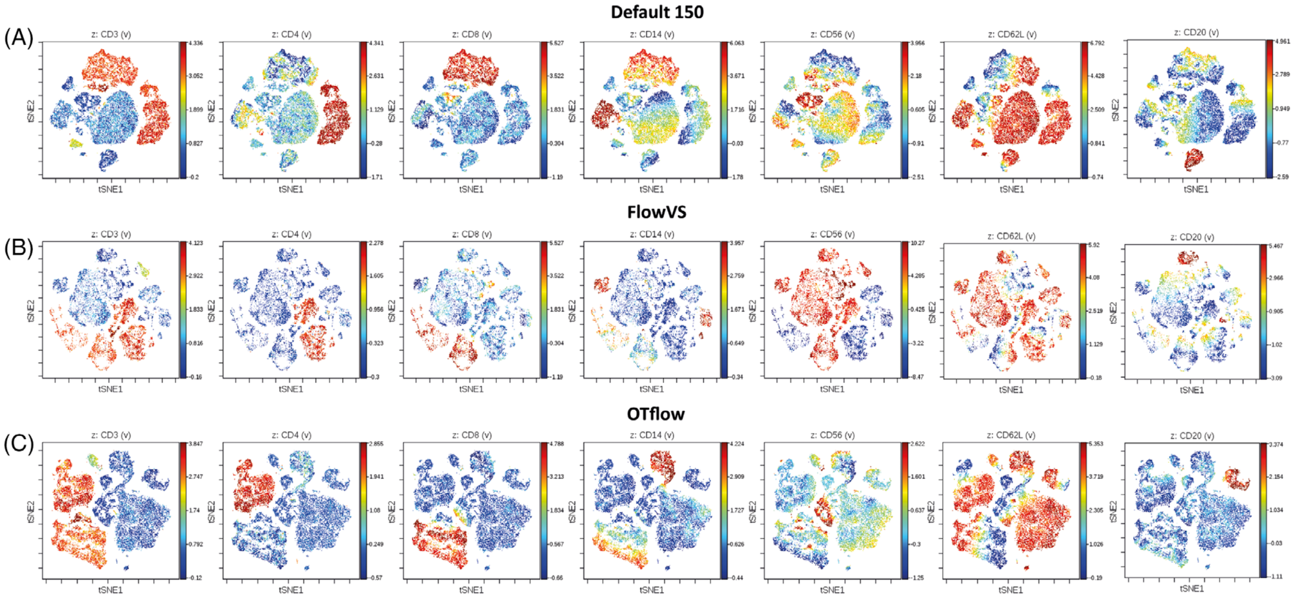

**FIGURE 4** viSNE maps of default transformed cofactor 150 (A), flowVS transformed (B) and OTflow transformed (C) HIV datasets. The colour coding of the cells represents the expression levels for the specific marker named in the title of the plots. The viSNE maps of the variously transformed datasets differ greatly from each other

minimum spanning tree or in a two dimensional grid. Cells with similar phenotypes are clustered together in the same node. For conciseness we only show the result of the flowVS and OTflow transformed HIV datasets, which differed the most from each other (Figure 5AB).

In the FlowSOM results of the flowVS transformed HIV dataset (Figure 5A) we identified branches containing nodes with similar phenotypes, only differing based on CD56 expression. Monocyte (CD14$^{\text{bright}}$), neutrophil (CD16$^{\text{bright}}$) and CD4$^{\text{bright}}$ and CD8$^{\text{bright}}$ T-cell (CD3$^{\text{bright}}$) subpopulations were divided into CD56$^{\text{low}}$ and CD56$^{\text{bright}}$ nodes (Figure 5C). As an example, the putative CD4$^{\text{bright}}$ and CD8$^{\text{bright}}$ T-cell subpopulations show differential expression of (mainly) marker CD56 due to peak splitting. This is in contrast to the OTflow

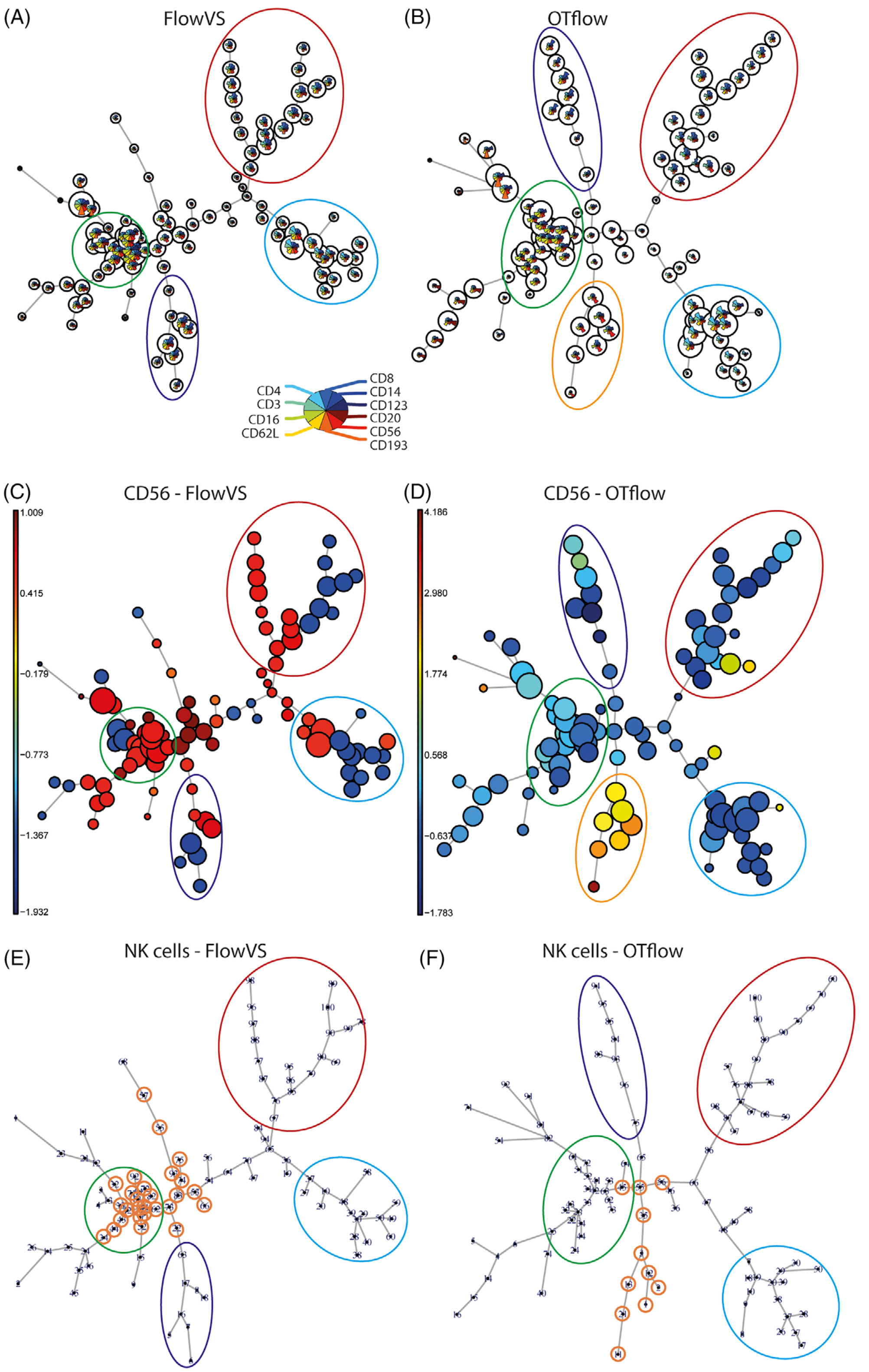

**FIGURE 5** (A–F) Leukocyte populations in flowVS transformed flowSOM map are divided over CD56bright and CD56low nodes. (A) flowSOM results of HIV data transformed with flowVS-defined cofactors. The mean marker expressions are visualized in each node by a pie chart. The size of the nodes corresponds to the relative amount of cells in each node. Some nodes of interest have been encircled with ellipses. Dark blue encircled nodes comprise CD14bright monocytes. Green encircled nodes comprise neutrophils (CD16brightCD62Lbright). Red encircled nodes comprise CD3brightCD8bright T cells. Light blue encircled nodes comprise CD3brightCD4bright T cells; (B) flowSOM minimum spanning tree built on the OTflow transformed HIV data. Nodes containing cells corresponding to the selected nodes in A have been encircled with the same colour coding. Dark blue encircled nodes comprise CD14bright monocytes. Green encircled nodes comprise neutrophils (CD16brightCD62Lbright). Red encircled nodes comprise CD3 brightCD8bright T cells Light blue encircles nodes comprise CD3brightCD4bright T cells. The orange encircled nodes comprise CD56brightCD16dim NK cells. (C–D) CD56 expression levels projected onto the flowSOM map of flowVS (C) and OTflow (D) transformed HIV data. The encircled nodes correspond to the encircled nodes in A and B. (E,F) Nodes to which NK cells were assigned are highlighted by orange circles in the flowVS (E) and OTflow transformed (F) flowSOM maps

transformed flowSOM map, where these cell populations all have low CD56 expressions (Figure 5D). The seven nodes with a CD56$^{bright}$ phenotype (Figure 5D, orange circle) in the OTflow map are also CD16$^{dim}$, fitting with the NK cell phenotype (Figure 5B, orange circle). When peak splitting occurs due to suboptimal transformation with poorly-defined cofactors, the appearing negative cell populations may occupy unique nodes at the expense of biologically distinct cell populations which may then be clustered together in the same node. This is indeed what we found. In the flowVS transformed flowSOM map CD56$^{bright}$CD16$^{dim}$ NK cells were not easily identified. We manually gated CD56$^{bright}$CD16$^{dim}$ NK cells in our flowSOM analysis to retrieve the flowSOM cluster numbers to which the cells were assigned. NK cells were not clustered together in a separate branch like the other cell types, but they were found back in the main central branch, overlapping with nodes which mainly occupied neutrophils (Figure 5E). This is in contrast to OTflow transformed flowSOM map, where the gated NK cells were all assigned to nodes in the same branch or to the nodes in the closest proximity of this branch. From this, we can conclude that suboptimal transformation of flow cytometry data, leading to peak splitting, results into less interperable flowSOM results which may lead to false conclusions.

### 3.3.4 | Citrus on the transformed HIV dataset

Citrus analysis [10], which is a classification method, was performed to identify which cell population-specific features were found to be discriminant for certain cell populations in the dataset were specific for healthy control individuals or for HIV patients. The results can be found in Appendix S1. Strikingly, we found different results when applying the method on the differently transformed datasets.

## 4 | RESULTS FROM OTHER DATASETS

Two additional datasets were used to show the effect of suboptimal transformations, one originating from a Lean vs Obese study and one from a Tuberculosis (TBC) study. The results are shown in the Supplementary Material IV. The PCA results of the suboptimal flowVS transformed lean vs obese dataset caused peak splitting, which resulted in the rise of nonexistent cell populations in a multidimensional analysis, potentially leading to misinterpretations. When OTflow was performed on the same dataset, the nonexistent cell populations were not present in the PCA results. Secondly, suboptimal transformation was shown to affect the outcome of viSNE results for the TBC dataset, producing artifacts due to peak splitting and, at least as important, hiding immunologically relevant information.

## 5 | SIMULATION STUDY RESULTS

A simulation study was performed with univariate data to visualize and quantify the behavior of the OTflow algorithm, and compare it with flowVS. See Supplementary Material V. From this study can be concluded that the OTflow algorithm performs well for all scenario's tested. In general, transformation of flow cytometry data might lead to relative deviations from reality. However, a shift of peaks as might be imposed by OTflow, does not lead to misleading biological conclusions. OTflow can deal with negative values in the data, and does not lead to peak splitting. This is in contrast to flowVS, for which negative values in the data may result in peak splitting, leading to misinterpretations of the data.

## 6 | DISCUSSION

MFC data transformation, required before multivariate methods such as PCA, viSNE and flowSOM can be applied, is often performed by bi-exponential transformations, such as arcsinh and logicle. We demonstrate that bi-exponential transformations can introduce artifacts in multivariate analysis results. An optimal display of the MFC data in both univariate and multivariate data representation greatly relies on the choice of the transformation parameter(s).

A cofactor of 150 for the arcsinh function, suggested by default in various methods and in Cytobank, rarely leads to proper representation of the cell populations in the data for all the markers. A disadvantage of the use of a single co-factor (of for instance 150) consists of the fact that different markers may necessitate different cofactors because the fluorescence intensities depend on the dye used, on the background fluorescence per dye, on marker expression, and on the compensation used to correct for spill-over. Notably, all these factors may vary per channel. Visual inspection of the transformed data per channel (not performed here) would be highly subjective, time-consuming and inefficient, especially with large panels.

The novel OTflow algorithm enables an automated, data-driven and validated estimation of channel-specific optimal parameters for both arcsinh and logicle transformations. As the number of parameters measured per cell continues to increase with the development of new flow cytometers [39], automated data-driven transformation becomes crucial. OTflow avoids artifacts such as merging two phenotypically different cell populations into one population by condensing two separate peaks into one spiky peak. OTflow also accurately prevents artifacts at low intensities by disregarding cofactors which lead to over-dispersion of the leftmost peak by splitting it in two cell populations (i.e. 'peak splitting'). This key feature of the algorithm represents one of the innovative and essential steps of OTflow compared to other methods developed to estimate transformation parameters such as flowVS and the calculation for the *W* parameter for the logicle function [15, 18].

As a result, OTflow outperforms methods such as flowVS and default-cofactor arcsinh transformation in properly displaying cell subpopulations and subsequent multivariate analyses. We showed with various MFC datasets (as shown for the HIV dataset in the main text and for two additional datasets in Supplementary Material IV) that unintentional peak splitting in one or more channels caused by flowVS defined cofactors or default cofactors dramatically propagates to higher dimensions when applying multivariate analysis methods. Moreover, these artifacts are independent to the type of MVA method used as it affects PCA, viSNE, flowSOM and Citrus results. For instance, in the HIV data, homogenous CD4$^{bright}$CD56$^{low}$ and CD8$^{bright}$CD56$^{low}$ T-cell populations showed differential expression in CD56 in the PCA space and they were over-fragmented in the viSNE map, due to transformation-induced peak splitting of CD56$^{low}$ cells, as confirmed by backgating. Suboptimal transformation greatly complicates multivariate analysis because the phenotypically identical populations are separated into multiple seemingly phenotypically different populations. This generates confusion and it is highly detrimental to subsequent immunological interpretation and conclusions. Also, we showed how the additional variability introduced by peak splitting may overshadow the more subtle but relevant biological information in the data. This was the case in PCA as well as flowSOM. In the flowSOM map peak splitting caused homogeneous cell populations to be divided over multiple nodes at the expense of 'real' phenotypes that were clustered together in one node. Citrus also indicated the negative peaks as most discriminant between the two groups considered. Next to the results from the HIV analysis, also the results from TBC analysis (Supplementary material IV) exemplify that peak splitting causes both the gain of false information in a dataset and the loss of immunologically relevant information.

We demonstrated that OTflow also aids in finding the optimal value for the *W* parameter for logicle transformation (Figure S12). The calculation of the optimal *W* as proposed by Parks et al [15] may not always lead to proper transformation of all the channels. Also here artifacts may be introduced, because the *W* is highly dependent on the negative intensities in the data. Secondly, because it focuses only on the first peak with the lowest expression, logicle transformation may not stabilize variance between the other cell populations with higher expressions.

For a widespread application of OTflow, some assumptions considered in the algorithm's steps are here discussed. The algorithm works by identifying uni-dimensional peaks in the iteratively transformed channels. By using a very sensitive function, OTflow can also detect very small peaks. However, some threshold in 'finding' peaks has to be imposed (Figure 1, Step 3). This is done to circumvent the detection of too large numbers of peaks not necessarily having biological meaning but associated to noise. OTflow thus removes peaks containing <5 cells, corresponding to 0.5% of the subsample considered (Figure 1, Step 3a). Especially in the extreme ends of the fluorescence range these are likely to be associated to noise or machine artifacts and therefore not to be introduced in the variance stabilization step. In some studies, however, very small cell populations may be expected. In this case it may be advisable to reduce this limit to 0.01 or 0.02% to avoid loss of relevant information, with the associated drawback of reducing the computational speed of the algorithm.

## 7 | CONCLUSION

Transformation of MFC data is an essential step before applying many types of univariate and multidimensional data analysis method. Suboptimal transformation due to poorly defined parameters can highly complicate the analysis results by introducing artifacts in the data and thus lead to misleading interpretations. The algorithm OTflow, here introduced, is a robust method for automated data-driven bi-exponential transformation of flow cytometry data. Most importantly, OTflow prevents split peaks to arise in the data, thereby preventing directly the false interpretation of distinct cell populations and/or the indirect failure (due to domination by transformation-induced variance) to observe phenotypically different cell subsets.

The versatility of the algorithm allows its application to any type of MFC dataset for the estimation of parameters for both arcsinh and logicle transformation. Integration of the algorithm in MFC analysis methods will facilitate rapid optimized transformation of the data and minimize the risk of misinterpretation.

## AUTHOR CONTRIBUTIONS

**Rita Folcarelli:** Conceptualization; formal analysis; investigation; methodology; software; visualization; writing-original draft. **Selma van Staveren:** Conceptualization; formal analysis; investigation; methodology; resources; visualization; writing-original draft. **Gerjen Tinnevelt:** Conceptualization; formal analysis; investigation; methodology; software. **Emily Cadot:** Formal analysis; investigation; methodology; software. **Nienke Vrisekoop:** Conceptualization; funding acquisition; resources; supervision; writing-review & editing. **Lutgarde Buydens:** Funding acquisition; project administration; supervision; writing-review & editing. **Leo Koenderman:** Conceptualization; funding acquisition; resources; supervision; writing-review & editing. **Jeroen Jansen:** Conceptualization; funding acquisition; methodology; supervision; writing-review & editing. **Oscar van den Brink:** Conceptualization; methodology; project administration; supervision; writing-review & editing.

## PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/cyto.a.24491.

## ORCID

*Selma van Staveren* https://orcid.org/0000-0002-2531-209X

## REFERENCES

1. Perfetto SP, Chattopadhyay PK, Roederer M. Innovation: seventeen-colour flow cytometry: unravelling the immune system. Nat Rev Immunol. 2004;4:648–55. https://doi.org/10.1038/nri1416
2. Chattopadhyay P, Perfetto S, Gaylord B, Stall A, Duckett L, Hill J, et al. Toward 40+ parameter fluorescence flow cytometry. XXIX congress of the International Society for Advancement of cytometry. Ft. Lauderdale, FL.: International Society for Advancement of cytometry;2014. p. 215–216.
3. Saeys Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nat Rev Immunol. 2016;16:449–62. https://doi.org/10.1038/nri.2016.56
4. Amir ED, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat Biotechnol. 2013;31:545–52. https://doi.org/10.1038/nbt.2594
5. Jansen JJ, Hilvering B, van den Doel A, Pickkers P, Koenderman L, Buydens LMC, et al. FLOOD: FLow cytometric orthogonal orientation

for diagnosis. Chemom Intel Lab Syst. 2016;151(December):126–35. https://doi.org/10.1016/j.chemolab.2015.12.001

6. Tinnevelt GH, Kokla M, Hilvering B, van Staveren S, Folcarelli R, Xue L, et al. Novel data analysis method for multicolour flow cytometry links variability of multiple markers on single cells to a clinical phenotype. Sci Rep. 2017;7:1–11. https://doi.org/10.1038/s41598-017-05714-1
7. Folcarelli R, van Staveren S, Bouman R, Hilvering B, Tinnevelt GH, Postma G, et al. Automated flow cytometric identification of disease-specific cells by the ECLIPSE algorithm. Sci Rep. 2018;8:1–18. https://doi.org/10.1038/s41598-018-29367-w
8. Qiu P, Simonds EF, Bendall SC, Gibbs KD Jr, Bruggner RV, Linderman MD, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol. 2011;29:886–93. https://doi.org/10.1038/nbt.1991
9. Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. Cytom Part A. 2015;87:636–45. https://doi.org/10.1002/cyto.a.22625
10. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. Proc Natl Acad Sci. 2014;111:E2770–7. https://doi.org/10.1073/pnas.1408792111
11. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006;7. https://doi.org/10.1186/1471-2164-7-142
12. Verwer B. BD FACSDiVa option (white paper). Becton: Dickinson and Company; 2002. http://www.bdbiosciences.com/ds/is/others/23-6579.pdf
13. Herzenberg LA, Tung J, Moore WA, Herzenberg LA, Parks DR. Interpreting flow cytometry data: a guide for the perplexed. Nat Immunol. 2006;7:681–5. https://doi.org/10.1038/ni0706-681
14. Roederer M. Spectral compensation for flow cytometry: visualization artifacts, limitations, and caveats. Cytometry. 2001;45:194–205. http://www.ncbi.nlm.nih.gov/pubmed/11746088
15. Parks DR, Roederer M, Moore WA. A new "logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. Cytom Part A. 2006;69A:541–51. https://doi.org/10.1002/cyto.a.20258
16. Finak G, Perez J-M, Weng A, Gottardo R. Optimizing transformations for automated, high throughput analysis of flow cytometry data. BMC Bioinformatics. 2010;11:546. https://doi.org/10.1186/1471-2105-11-546
17. Kotecha N, Krutzik PO, Irish JM. Web-based analysis and publication of flow cytometry experiments. Curr Protoc Cytom. 2010;Chapter 10:Unit10.17. https://doi.org/10.1002/0471142956.cy1017s53
18. Azad A, Rajwa B, Pothen A. flowVS: channel-specific variance stabilization in flow cytometry. BMC Bioinformatics. 2016;17:1–14. https://doi.org/10.1186/s12859-016-1083-9
19. Tibshirani R. Estimating transformations for regression via additivity and variance stabilization. J Am Stat Assoc. 1988;83:394–405. https://doi.org/10.1080/01621459.1988.10478610
20. Bartlett MS. The square root transformation in analysis of variance. R Stat Soc. 1936;3:68–78. https://doi.org/10.2307/2983678
21. Ray S, Pyne S. A computational framework to emulate the human perspective in flow cytometric data analysis. PLoS One. 2012;7:e35693. https://doi.org/10.1371/journal.pone.0035693
22. Jarque CM, Bera AK. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ Lett. 1980;6:255–9. https://doi.org/10.1016/0165-1765(80)90024-5
23. Jolliffe IT. Principal component analysis. Principal Component Analysis. 2nd ed. New York: Springer; 2002. p. 1–405.
24. Box GEP, Cox DR. An analysis of transformations (with discussion). J R Stat Soc B. 1964;77:209. https://doi.org/10.2307/2287791
25. Botev ZI, Grotowski JF, Kroese DP. Kernel density estimation via diffusion. Ann Stat. 2010;38:2916–2957. https://doi.org/10.1214/10-AOS799
26. Find local maxima - MATLAB findpeaks - MathWorks Benelux. https://nl.mathworks.com/help/signal/ref/findpeaks.html. Accessed February 14, 2021.
27. Jarque CM, Bera AK. A test for normality of observations and regression residuals. Int Stat Rev. 1987;55:163. https://doi.org/10.2307/1403192
28. Lugli E, Pinti M, Nasi M, Troiano L, Ferraresi R, Mussi C, et al. Subject classification obtained by cluster analysis and principal component analysis applied to flow cytometric data. Cytom Part A. 2007;71A:334–44. https://doi.org/10.1002/cyto.a.20387
29. Costa ES, Pedreira CE, Barrena S, Lecrevisse Q, Flores J, Quijano S, et al. Automated pattern-guided principal component analysis vs expert-based immunophenotypic classification of B-cell chronic lymphoproliferative disorders: a step forward in the standardization of clinical immunophenotyping. Leukemia. 2010;24:1927–33. https://doi.org/10.1038/leu.2010.160
30. Tung JW, Parks DR, Moore WA, Herzenberg LA, Herzenberg LA. New approaches to fluorescence compensation and visualization of FACS data. Clin Immunol. 2004;110:277–83. https://doi.org/10.1016/j.clim.2003.11.016
31. Bagwell CB. HyperLog - A flexible log-like transform for negative, zero, and positive valued data. Cytom Part A. 2005;64:34–42. https://doi.org/10.1002/cyto.a.20114
32. Hahne F, LeMeur N, Brinkman RR, Ellis B, Haaland P, Sarkar D, et al. flowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinformatics. 2009;10:106. https://doi.org/10.1186/1471-2105-10-106
33. Bioconductor - flowCore. https://www.bioconductor.org/packages/release/bioc/html/flowCore.html. Accessed February 14, 2021.
34. Qian Y, Liu Y, Campbell J, Thomson E, Kong YM, Scheuermann RH. FCSTrans: an open source software system for FCS file conversion and data transformation. Cytom Part A. 2012;81:353–356. https://doi.org/10.1002/cyto.a.22037
35. Lemster BH, Michel JJ, Montag DT, Paat JJ, Studenski SA, Newman AB, et al. Induction of CD56 and TCR-independent activation of T cells with aging. J Immunol. 2008;180:1979–90. https://doi.org/10.4049/jimmunol.180.3.1979
36. Potter SJ, Lacabaratz C, Lambotte O, Perez-Patrigeon S, Vingert B ˆ, Sinet M, et al. Preserved central memory and activated effector memory CD4+ T-cell subsets in human immunodeficiency virus controllers: an ANRS EP36 study. J Virol. 2007;81:13904–15. https://doi.org/10.1128/JVI.01401-07
37. Vassena L, Giuliani E, Buonomini AR, Malagnino V, Andreoni M, Doria M. Brief report: L-selectin (CD62L) is downregulated on CD4+- and CD8+T lymphocytes of HIV-1-infected individuals naive for ART. J Acquir Immune Defic Syndr. 2016;72:492–7. https://doi.org/10.1097/QAI.0000000000000999
38. Kononchik J, Ireland J, Zou Z, Segura J, Holzapfel G, Chastain A, et al. HIV-1 targets L-selectin for adhesion and induces its shedding for viral release. Nat Commun. 2018;9:1–15. https://doi.org/10.1038/s41467-018-05197-2
39. Robinson JP, Roederer M. Flow cytometry strikes gold. Science. 2015;350:739–40. https://doi.org/10.1126/science.aad6770

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Folcarelli R, van Staveren S, Tinnevelt G, Cadot E, Vrisekoop N, Buydens L, et al. Transformation of multicolour flow cytometry data with OTflow prevents misleading multivariate analysis results and incorrect immunological conclusions. Cytometry. 2022;101:72–85. https://doi.org/10.1002/cyto.a.24491