Received: 22 July 2021 | Revised: 10 April 2022 | Accepted: 25 April 2022

DOI: 10.1002/cem.3402

SPECIAL ISSUE - RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

Comprehensive multivariate evaluation of the effects on cell phenotypes in multicolor flow cytometry data using ANOVA simultaneous component analysis

Carlo G. Bertinetto[1] | Roy Spijkerman[2,3,4] | Lillian Hesselink[2,3] | Gerjen H. Tinnevelt[1] | Coen C. W. G. Bongers[5] | Geert J. Postma[1] | Maria T. E. Hopman[5] | Leo Koenderman[3,4] | Jeroen J. Jansen[1]

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

[2]Department of Trauma Surgery, University Medical Center Utrecht, Utrecht, The Netherlands

[3]Center for Translational Immunology (CTI), University Medical Center Utrecht, Utrecht, The Netherlands

[4]Department of Respiratory Medicine, University Medical Center Utrecht, Utrecht, The Netherlands

[5]Department of Physiology, Radboud Institute for Health Sciences (RIHS), Radboud University Medical Center, Nijmegen, The Netherlands

Correspondence
Carlo G. Bertinetto, Institute for Molecules and Materials (Analytical Chemistry), Radboud University, Heyendaalseweg 135, 6525 AJ, Nijmegen, The Netherlands.
Email: cgbertinetto@gmail.com

Abstract

This work proposes an approach to assess the effects observed in multicolor flow cytometry (MFC) experiments, for all markers and experimental factors simultaneously. It achieves this end by extending ANOVA simultaneous component analysis (ASCA), a multivariate version of ANOVA, to flow cytometry data. It is based on an initial multiset PCA model to describe the main variation patterns of cell marker expression, followed by an ASCA model on the histograms built from these PCA scores. This approach allows for determining the variations in cell phenotype distribution that are related to the experimental design. On a data set from a study of the immune response to prolonged physical exercise, the proposed method computed the effect size and statistical significance of all the experimental factors and their interactions. Most notably, it provided easily interpretable submodels for the overall effect of the walking exercise and for the interaction between exercise and the responsiveness to a bacterial stimulus. The application of a time-guided sequential clustering algorithm to the ASCA scores revealed a stratification of the studied individuals based on their neutrophil activation dynamics. These effects were not clearly detectable using PCA alone. In comparison with pairwise classification models by DAMACY (a discriminant analysis method for MFC data), ASCA results were less detailed in describing differences between specific samples, but had the advantage of modeling several factors and levels simultaneously. Such characteristics make the proposed implementation of ASCA an effective and complementary addition to the chemometric methodologies for the analysis of MFC data.

KEYWORDS

Discriminant Analysis of MultiAspect Cytometry (DAMACY), experimental design, innate immune response, physical exercise, Roy

1 | INTRODUCTION

Multicolor Flow Cytometry (MFC) is one of the most established and insightful methods for single cell analysis,[1,2] with many applications in medical and environmental sciences. It is based on aligning the cells of a given sample, usually after appropriate staining, along a narrow funnel and analyzing them one by one, typically by recording their fluorescence from a laser beam. In such way, all measured cells are characterized by the expression of a certain number of markers, normally 5–20 depending on the instrument. This high-throughput technique enables a very detailed description of the cell variability, allowing detection of pathologies and support identification of mechanisms underlying biological processes.[3] Recently, its clinical versatility has been further enhanced by the development of compact, (semi) automated MFC systems[4] that can analyze samples in situ. This greatly extends the employability of MFC from the laboratory to *point-of-care* diagnostics.

The large volume of data generated by MFC, from thousands to millions of cells per sample, is highly information-rich on the immune system. The measurements on many different cells within the same sample may identify cells with differential expressions of immune markers, either associated with emergence of cell types with complementary function in the immune system or with activation of the functionality in cells that were already present. This in turn represents changes in the entire sample under scrutiny that may indicate system-wide immunological changes of potential diagnostic value.

Nevertheless, this wealth of data poses considerable challenges to the analysis task. Conventionally, such data is analyzed with a multi-gating approach, restricting the analysis to two variables at a time. This is still widely used in clinical settings but becomes resource-intensive and time-consuming even with a low number of markers, compared to the emerging potential of automated, digital approaches. Manual gating strategies are furthermore strongly biased, as they rely strongly on previous interpretations and thus easily miss emerging, unknown patterns that do not align with the chosen strategy.

Various computational methods have been developed to perform a more efficient and less biased analysis.[5] Some of the most widely used are based on non-linear methods, such as stochastic neighbor embedding,[6,7] self-organizing maps,[8] density-normalized clustering,[9] or deep convolutional neural networks.[10] These methods are often accurate and efficient for tasks of cell classification and in some cases even cell sorting, but they do not allow for unambiguous interpretation of the biological differences among the resulting different types of cells, as the results of these methods are non-deterministic. Therefore, additional specific analysis for each subgroup is indispensable.

Deterministic linear models, such as Principal Component Analysis (PCA) or Partial Least Squares (PLS), provide this kind of information. Although they have been criticized for failing to grasp the complex non-linearities associated to changes in the immune system, the transparency of their results in terms of the immune markers measured on every cell is indisputably superior, because their loadings point explicitly to the direction of biological variability.[11] Following this approach, several methods dedicated to MFC data have been developed, such as DAMACY,[12] FLOOD[13] and ECLIPSE.[14] They all are based on a combination of models applied sequentially: an initial multiset PCA[15] to provide a compact description of the distributions of marker values throughout the analyzed samples, followed by other models to relate these distributions to the problem of interest, usually a discrimination between control and response cases. They have been applied successfully to data from clinical studies analyzing the immune responses to, for example, experimental endotoxemia,[12–14,16] hematological cancers[12] and asthma.[12,14]

These methods have also been applied to non-dichotomous cases such as time series, notably in an investigation using FLOOD of the effect of several consecutive days of endurance exercise.[17] This was done by building a PCA model with the samples from the first time point, thus treated as a control case, and projecting all the remaining samples on that (control) PCA space. The results revealed the mobilization of certain cell phenotypes associated with systemic inflammation and immunosuppression. However, this mainly unsupervised method often struggles to detect small effects that are hidden by larger sources of variation such as differences among individuals. Furthermore, it does not provide a direct and quantitative estimation of several effects simultaneously. Both these goals are attainable if the information on the experimental design, for example, time points, case groups and any other factor, is incorporated into the data analysis model.

ANOVA Simultaneous Component Analysis (ASCA)[18,19] is one of the leading methods to perform this task. It consists in a multivariate version of ANOVA that allows for identifying the multivariate responses associated to specific factors and interactions, as well as determining their statistical significance. It has been employed for a wide range of data from spectroscopy[20–22] to –omics,[23–26] but not in flow cytometry, except for a study in which it was used only on the main cell counts.[27] Therefore, the aim of the current work is to propose an implementation of ASCA for MFC that takes

into account the complexity of this type of data. The advantage of such data analysis tool in medical and biological research would be to enable a comprehensive and simultaneous evaluation of all the effects of interest on all markers, without needing the laborious task of examining specific factors or subsets of markers at a time.

The proposed implementation is devised by extending DAMACY,[12] which is a method to perform discriminant analysis on MFC data. Moreover, we will demonstrate through a case study how the ASCA results can be integrated with those of complementary chemometric methods to obtain a more comprehensive analysis approach. These methods also include an adapted clustering approach that operates following the chronological order of the data and is thus particularly useful to distinguish the main time patterns.

The data under consideration comes from a study on the response of the innate immune system to prolonged physical exercise.[28] This is a topic of great interest[29,30] because such response is not fully understood and can sometimes be harmful, particularly on elderly subjects that are more prone to sub-optimal immune responses resulting in inflammation and/or insufficient protection.[31] Moreover, because several clinical conditions affect this physiological outcome, diagnostic tools that rely on the immune response at rest may be improved by observing the response in a challenged state.[32] A more informative chemometric analysis may, thus, help find the exercise load that optimizes the greatest health benefits, as well as biomarkers that enable targeted monitoring of the relevant health conditions.[33–35]

2 | MATERIALS AND METHODS

2.1 | Study design and measurements

The study cohort consisted of 45 subjects (aged 64 ± 6.8 years, 35% female) who participated to the 2018 edition of the 4-Day Marches (http://www.4daagse.nl/en/), a large walking event that takes place every July in Nijmegen, the Netherlands. Every participant walked an assigned distance (30, 40, or 50 km, depending on the age and physical conditions) at a self-selected pace, every day for four consecutive days. The cohort was also subdivided into three groups: healthy controls, users of statin (a commonly prescribed medication to lower blood cholesterol levels[36]) without side effects of muscle complaints (here referred to as Statin 1) and statin users with muscle complaints (Statin 2), containing 15, 14 and 16 individuals, respectively. A more detailed description of this cohort is given elsewhere.[37]

All subjects were measured on-site at baseline (1 or 2 days before the event) and immediately after every day of walking for the first 3 days, by taking a blood sample and loading it into an automated AQUIOS CL® "Load & Go" flow cytometer (Beckman Coulter, Miami, FL, USA).[4] Each sample was split into two aliquots, of which one was analyzed directly, whereas the other was treated with *N*-formyl-methionyl-leucyl-phenylalanine (fMLF), a chemotactic factor commonly used to stimulate the innate immune system by mimicking the oligopeptides released by bacteria.[38] The two types of aliquots are here referred to as fMLF− and fMLF+, respectively. Subsequently, all samples were stained with an antibody mix containing CD16, CD62L, CD35, CD10 and CD11b, as described in detail in a previous work.[37] They were measured by a fully automated flow cytometer, AQUIOS CL® (Beckman Coulter Life Sciences, Miami, FL, USA),

TABLE 1 Experimental design of this study

Factor	No. of levels	Level names
Individuals	45	1-45
Group	3	Healthy controls Statin users 1 (no side effects) Statin users 2 (with side effects)
Time point	4	Baseline Day 1 Day 2 Day 3
Bacterial stimulation	2	fMLF− fMLF+

Note: Besides *individuals*, all factors are crossed.

which has a 488 nm diode laser, two light scatter channels (forward scatter and side scatter), five fluorescence channels and an electronic volume measure. A scheme of the experimental design is shown in Table 1.

This study was approved by the Medical Ethical Committee of the Radboud University Medical Center under protocol number CMO-nr 2007-148 and all participants gave written informed consent before participation. All procedures performed in this study were in accordance with the 1964 Helsinki declaration and its later amendments.

2.2 | Chemometric analysis

2.2.1 | Preprocessing and *base*PCA

The flow cytometry measurements were acquired by the AQUIOS CL® instrument as .lmd files and exported into FlowJo® analysis software (Tree Star Inc., Ashland, OG, USA). The data were gated for neutrophils, the most numerous cell type of the innate immune system, based on the expression of CD16 and CD62L[39]; this gating was checked by two independent researchers. The final data structure consisted of 342 samples, containing 10,000–40,000 cells each, and measured over six variables, that is, the five above-mentioned antibody markers and Forward Scatter (FS). These data were analyzed according to the scheme depicted in Figure 1. The preprocessing consisted of several operations, beginning with compensating for the spillover among different fluorescence channels, and transforming using arcsinh. To avoid artifacts caused by values much lower than −1, prior to arcsinh transformation each column of the whole data was divided by a cofactor, which was initially chosen as the absolute value of the 99th percentile of all the negative values for that column. This cofactor was then checked by visual inspection of the resulting histogram, and modified if the latter deviated considerably from a Gaussian distribution or if spurious negative peaks were observed.[40] The arcsinh-transformed data were divided by the number of cells in each sample, then median-centered and scaled as done in a previous paper.[15] Depending on which method was used afterwards, the centering and scaling was executed either over the whole data set or over the sample that was identified as reference/control (see below).

The preprocessed data of all the concatenated samples were first modeled using simultaneous component analysis (SCA),[41] which is a family of methods for the analysis of data arrays with a shared mode. Its simplest form, also known as SCA-P[42] and employed here, consists in a multiset expansion of PCA to populations measured with a common set of variables; hence, the SCA model built on the preprocessed data is here referred to as *base*PCA. After projecting the samples onto the first two (*base*)PCs, a 2D-histogram (100 × 100 bins) of each one was calculated on the PCA scores as described before.[12] These histograms, a few examples of which are shown in Figure 2, provide an overview of the multivariate information of all cells measured in each sample. The number of bins should be chosen so that the distribution is sufficiently representative but also described with sufficient detail. As a rule of thumb, this number should be equal or lower than the number of cells. All *base*PCA histograms were smoothed and normalized to the unit sum of all bins, then unfolded and concatenated into a single matrix of dimension $N \times 10{,}000$, with N the number of samples considered in each case: some models were built on the whole data set, others on specific subsets only, see Section 3. The

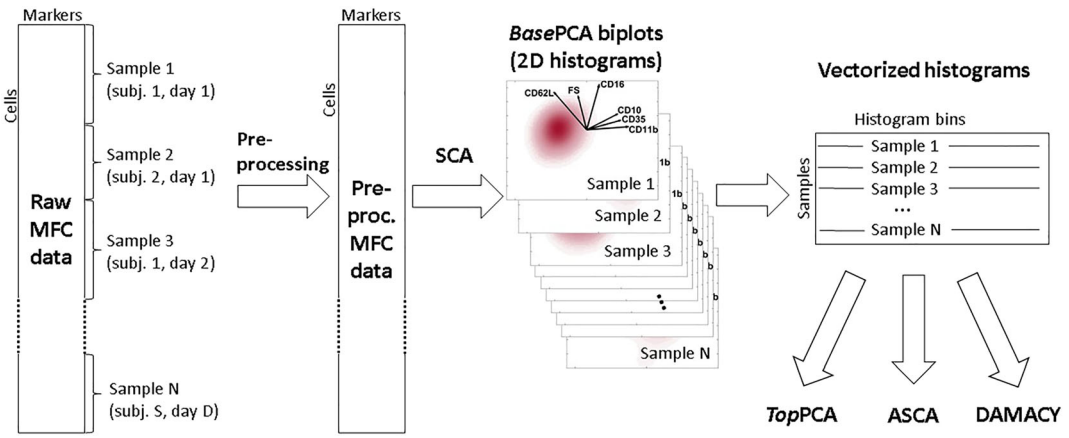

FIGURE 1 Scheme of the data analysis performed in this study. For the meaning of the acronyms (SCA, *base*PCA, *top*PCA, ASCA and DAMACY) see main text. The 2D histograms can also be generalized to histograms of several dimensions (see Section 3.1)

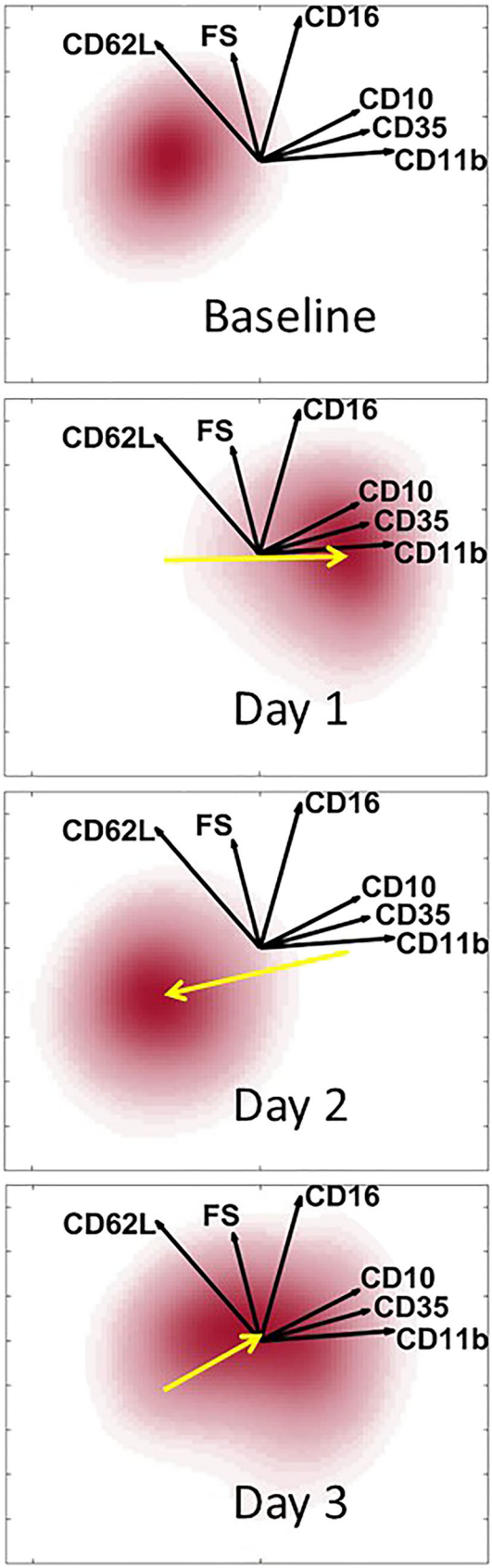

FIGURE 2 2D-histograms of *base*PCA scores for subject 14, fMLF−, on each measured day. The red spots depict the cell density with respect to the markers indicated by the loadings (black arrows); the yellow arrows indicate the shift of the cell density peak as compared to the previous time-point

resulting matrix was used as input for three further chemometric methods. The first was another PCA model (here denoted as *top*PCA), which was calculated in the standard way after mean-centering the matrix of unfolded 2D-histograms. The other methods were ANOVA Simultaneous Component Analysis (ASCA) and Discriminant Analysis of Multi-Aspect CYtometry (DAMACY).

2.2.2 | ANOVA simultaneous component analysis (ASCA)

ASCA is a multivariate approach to ANOVA, able to highlight the patterns that correlate with different factors or interactions of an experimental design.[18,24] The method consists of two main steps: (i) decomposition of the data matrix into a sum of matrices that explain each effect, and (ii) dimensional reduction of each effect matrix by means of PCA (or, more specifically, simultaneous component analysis[43]). The data decomposition used here was:

$$\mathbf{X} = \mathbf{X_m} + \mathbf{X_d} + \mathbf{X_s} + \mathbf{X_f} + \mathbf{X_{ds}} + \mathbf{X_{df}} + \mathbf{X_{sf}} + \mathbf{X'_e}$$

$$\mathbf{X'_e} = \mathbf{X_i} + \mathbf{X''_e} \tag{1}$$

$\mathbf{X}$ is the matrix of vectorized 2D-histograms of the *base*PCA model, built after median-centering and scaling over the data set that is being modeled. $\mathbf{X_m}$ is a matrix with N identical rows containing the column means of $\mathbf{X}$; $\mathbf{X_d}$, $\mathbf{X_s}$ and $\mathbf{X_f}$ are the effect matrices of the factors (*walking*) *day*, *statin group* and *fMLF*, respectively; $\mathbf{X_{ds}}$, $\mathbf{X_{df}}$ and $\mathbf{X_{sf}}$ are the effect matrices of their binary interactions, $\mathbf{X'_e}$ is the residual of this first decomposition, which is in turn decomposed into $\mathbf{X_i}$ (i.e., the effect matrix for the factor *individual*, nested in the *statin* factor) and the final residuals $\mathbf{X''_e}$. The final residuals also include the interactions of *individual* with *day* and with *fMLF*, as normally done for random factors.[44,45] Because ASCA does not perform an explicit assessment of random factors, the latter are accounted for only by how they affect the spread and statistical significance of the fixed factors. For suggestions on how to include the explicit treatment of random factors, see discussion in Section 3.2.3. This decomposition does not include quadratic terms, as commonly done in ASCA and other multivariate ANOVA methods. The unbalancedness of the design was handled by applying type III sum-of-squares corrections,[46] which calculate each effect after removing the portion that is explained by the other factors and interactions. As demonstrated by Thiel et al.,[46] the consequence of not accounting for design unbalancedness may be an overestimation of some of the effects, especially interactions. The PCA models calculated on each effect matrix are referred to as ASCA submodels. The significance of each factor and interaction was tested using the sum-of-squares of the relevant effect matrix and comparing it with the null distribution generated by 500 permutations.[47] Additional ASCA models were built on fMLF− or fMLF+ samples only, after removing the *fMLF* factor and its relevant interactions from (1). Certain submodels were also calculated on effect matrices combining a factor with one of its interactions (e.g., $\mathbf{X_{f+df}} = \mathbf{X_f} + \mathbf{X_{df}}$), applying type III sum-of-squares corrections in this case, as well.

2.2.3 | Other employed methods: DAMACY and sequential clustering

DAMACY is an implementation of Orthogonal Partial Least Squares-Discriminant Analysis (OPLS-DA)[48] for flow cytometry data[12] and it was employed here to describe the local contrasts between specific days in more detail than ASCA. In particular, the constructed models discriminated between each walking day and the next, as well as between baseline and any other day. The median-centering and scaling was performed in a paired manner, using the median and median absolute deviation of the sample from the earlier of the two considered days. The prediction scores, obtained from a double cross-validation taking the pairs into account, indicates how well each sample can be distinguished between the days, and the weight vector, refolded into a 100 × 100 matrix, points to which cells in the 2D histograms are more represented on which day. A p-value for the predicted classification was calculated by means of a permutation test (100 permutations).

A cluster analysis was performed on the matrix $\mathbf{T}$ of individual time-trajectories, built by collecting the first two scores of the ASCA submodel that combines the factor *individual* with the *day-individual* interaction. The scores of each subject were arranged on the same row in chronological order (PC1 baseline, PC2 baseline, PC1 day 1, PC2 day 1, ...), for a final dimension of $n \times 8$, with n the number of individuals. Missing values were imputed with the Missing Data

Imputation Toolbox[49] using the default options, that is, Trimmed Scores Regression, 5000 iterations, a tolerance of 10^{-10} and 3 PCs as determined from cross validation. Subjects were clustered by applying agglomerative hierarchical clustering[50] on **T**, with Euclidean distance and Average linkage, in a sequential way. In particular, initial clusters were obtained from the distances calculated on the baseline scores (first two columns of **T**); each of these clusters were then subdivided according to the scores of day 1 (third and fourth column of **T**), and the same operation was repeated in nested fashion for days 2 and 3. This sequential procedure allows for incorporating information on the chronological order of the measurements. The cutoff distance used to determine the number of clusters, decided upon visual inspection of the dendrograms, was set as 0.02 (larger than all the 95% confidence intervals obtained by bootstrapping). Four isolated data points were merged with their nearest cluster.

All calculations were performed in Matlab R2018b (The Mathworks, USA).

3 | RESULTS AND DISCUSSION

Of the 45 measured subjects, three dropped out after 1 day of walking, two dropped out after 2 days and one, despite completing the event, did not show up for analysis at day 2. One subject (of those who dropped out after 2 days) turned out to be CD16-receptor deficient, which made him unsuitable to the analysis. After excluding this subject, the remaining data from 44 participants set comprised 336 samples, of which baseline and day 1 had both 88, whereas days 2 and 3 had both 80, always split exactly in half between fMLF− and fMLF+.

3.1 | PCA reveals only the largest patterns

The *base*PCA model built on the whole data set explained 58% and 19% of the total variance for the first and second PC, respectively. These two PCs also correspond to a minimum of the Predicted Residuals Sum of Squares (PRESS), as determined by a 10-fold cross validation. The *base*PCA loadings, depicted as the arrows in several plots, including Figure 2, indicate that CD10, CD11b and CD35 are strongly correlated, which is expected because they are all markers of neutrophil activation.[4] CD16 and CD62L, which are markers related to respectively cell maturity and adhesion, are instead more correlated to FS, which is a measure of the cell size and appears orthogonal to the activation markers. The *base*PCA 2D-histograms are an efficient way to represent all the major multivariate information of flow cytometry data in a single plot.[13] Here, they allow for an easy visualization of the changes occurring between different samples. For instance, the histograms in Figure 2 are a representation of the neutrophil phenotype of subject no. 14 during the measured days. The most apparent changes are shifts in the position of the distribution peak, indicating a different average phenotype of the cell population; a slight elongation is also observed on day 3. Even more detailed data representations would be obtained by using more than two *base*PCs, and the method described in this paper would remain essentially the same, apart from the higher memory requirements to compute a high-dimensional histogram (unless the bin resolution is reduced). However, such histogram would bring visualization challenges that are not in the scope of this paper.

On the matrix of the (unfolded) 2D-histograms of the *base*PCA, a second PCA model was constructed to highlight the main variations among the cell distributions throughout the data set. The loadings and scores from PCs 1–4 of this model, referred to as *top*PCA, are shown in Figure 3A. The loadings of the unfolded histograms are refolded into 2D-histograms, with red and blue areas indicating positive and negative density with respect to the mean (as the histogram data are first mean-centered). The loadings of the *base*PCA are projected on the same plot as arrows. This superposition is valid because the variables of the *top*PCA loadings are the same as in the *base*PCA 2D-histograms, that is, the histogram bins; the same consideration is true for the loadings or regression coefficients of other models presented below (ASCA and DAMACY, see Section 3.2.1). As can be seen from the loading histograms, each PC describes a specific effect: PC1 corresponds to a shift of the cell density towards higher values of CD10, CD11b and CD35, that is, an increase in neutrophil activation; PC2 corresponds to a spreading of the cell density towards CD62L as well as the activation markers; PC3 and following indicate more complex deformations of the cell density.

The first two *top*PCA scores, shown in Figure 3B, are plotted separately for each day, using dots for fMLF− and crosses for fMLF+. Each point is also characterized with a color and a number indicating the statin group and the subject ID, respectively. Because the fMLF stimulation causes a strong neutrophil activation, all the fMLF+ samples appear on the right side of the plot. However, a few fMLF− samples show a similar level of activation, especially on days 1 and

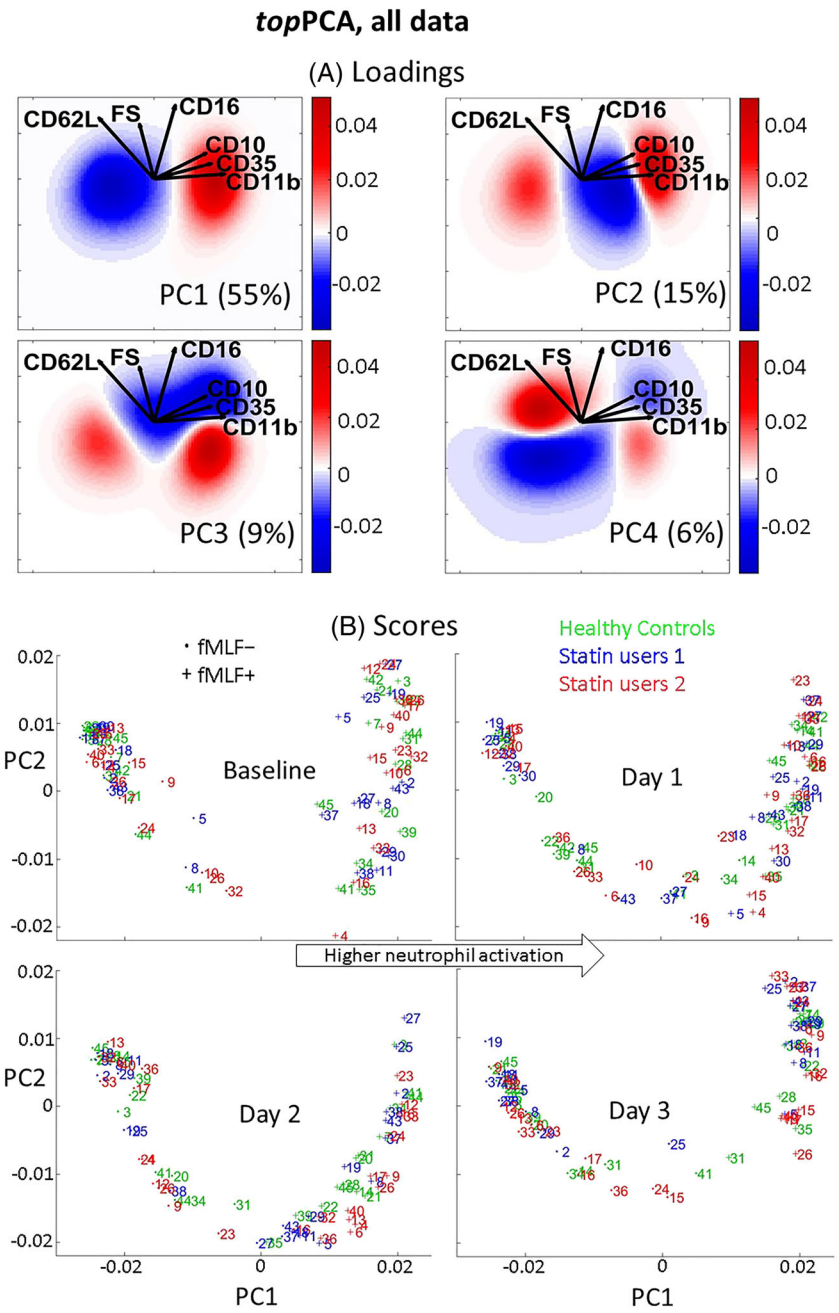

FIGURE 3 *top*PCA model on the full data set. (A) Loadings of PCs 1–4 (percentage of explained variance in parentheses); the arrows refer to the same *base*PCA loadings as Figure 2. (B) Scores of PCs 1–2 of each time-point separately; dots and crosses indicate fMLF− and fMLF+ samples, respectively; the numbers indicate the subject ID, whereas the colors indicate the statin group (see legend). The axes are identical for all plots

2. From these plots we cannot detect any clear pattern related to the time sequence, nor to the statin groups. Moreover, the horseshoe distributions observable in every plot raise the suspicion of representation artifacts.[51]

Subsequently, we built a *base*- and *top*PCA model on the fMLF− data only, whose loading and score plots are shown in Figure 4. This partial model removes the largest effect observed in the data, that is, the fMLF stimulation, and thus may facilitate the detection of patterns caused by smaller effects. The first two PCs of this *base*PCA explain 48% and 21% of the fMLF− data variance, respectively. Just like the full model, the loading of PC1 (Figure 4A) describes neutrophil activation, but PC2 instead describes a shift in cell density aligned towards higher CD16, corresponding to an increase in cell age. PC3 describes a spreading of the cell density (similar to PC2 of the full model), whereas PC4 and following are related to more complex effects.

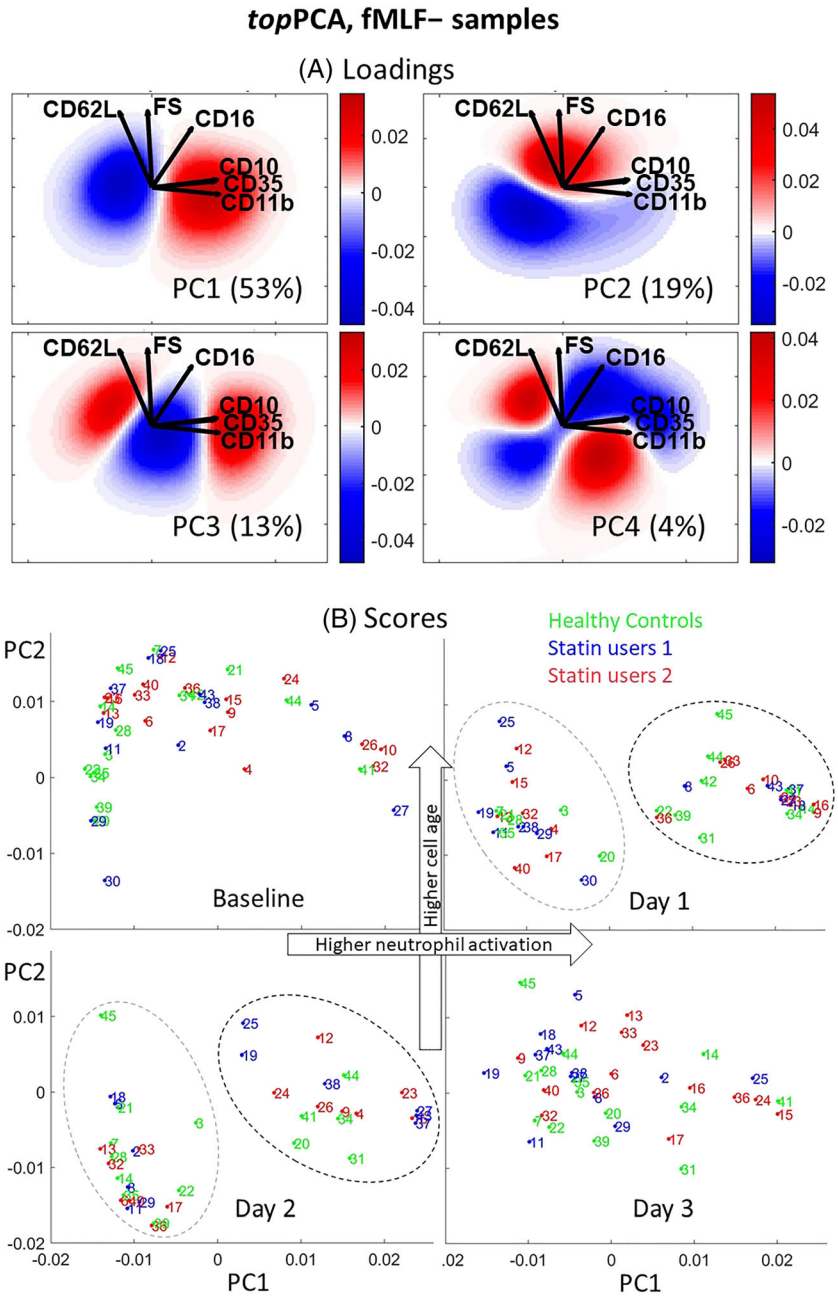

FIGURE 4 *top*PCA model on fMLF− samples. (A) Loadings of PCs 1–4 (percentage of explained variance in parentheses); the arrows refer to the *base*PCA loadings obtained from the fMLF− data. (B) Scores of PCs 1–2 of each time-point separately; the numbers indicate the subject ID, whereas the colors indicate the statin group (see legend). The axes are identical for all plots. On days 1 and 2, the subjects cluster into an activated and non-activated group, encircled by the dashed lines in black and gray color, respectively

The score plots (Figure 4B) do not show any evident horseshoe distribution anymore, and we can also find some notable patterns and groupings. At baseline, the subjects are spread rather evenly along PC1, spanning different degrees of neutrophil activation. On days 1 and 2, the individuals cluster into two main groups of "activated" and "non-activated" subjects. Interestingly, the individuals in each cluster are not the same between the 2 days, an observation that points to different patterns of activation-deactivation throughout the data set. The average PC2 values are lower on days 1 and 2 than baseline, especially for the non-activated groups. On day 3, the values of both PC1 and PC2 tend to shift towards the middle of the range, and the two clusters approximately merge back into a single one. Nevertheless, it is still not clear what patterns, if any, are related to the days of walking or the statin grouping.

TABLE 2 Explained variances (%) of the effects analyzed by the ASCA models used in this work

	% explained variance[a]		
Factor or interaction	**Full model**	**Partial (fMLF−)**	**Partial (fMLF+)**
fMLF	**47.4**		
Day	**5.0**	**12.0**	**21.6**
Statin	0.39	1.8	1.2
Individual	**9.6**	**36.6**	**34.6**
fMLF-day	**4.4**		
fMLF-statin	**0.4**		
Day-statin	0.8	1.6	2.2
Residual	32.9	47.8	40.4

Note: Those that are marked in bold are significant according to the permutation test ($p < 0.01$).
[a]Within the specific data set used for the model.

3.2 | ASCA reveals all patterns associated to the experimental design

The explained variances and statistical significance of the effects for all ASCA models calculated in this paper are presented in Table 2. The model built on the full data set is dominated by the *fMLF* effect, with more than 47% of the explained variance. As expected, the corresponding submodel has a loading (not shown) very similar to the *top*PC1 in Figure 3, and simply describes the neutrophil activation occurring in the fMLF+ samples. However, ASCA also finds other effects that are statistically significant (i.e., at least one group/level is significantly different from the others) from the *day* and *individual* factors, as well as the *fMLF-day* and *fMLF-statin* interactions. A rather large portion of the variance is left in the residuals, which contain the individual deviations to the average *fMLF* and *day* effects.

3.2.1 | Effect of *day* factor

Figure 5A shows the scores (with the projected residuals) and loadings of the *day* submodel, explaining 5% of the total variance. The black arrows highlight the average trajectory of the time points: from baseline to day 2 the trend is towards the left, where from day 2 to 3 it points approximately in the opposite direction. However, the loadings are not easy to interpret, because at least two shifts in cellular density are present, most likely originating from the fMLF− and fMLF+ samples, respectively. These density regions are tentatively encircled with green continuous and dashed lines, but because of a considerable overlap (especially in the blue areas) we cannot determine with certainty which density shift is related to which type of samples.

Therefore, it is convenient to build also ASCA models on half of the data (either fMLF− or fMLF+), analogously to what we did for the *top*PCA. For the model on fMLF− samples, the *day* submodel explains only 12% of the variance (of fMLF− data), of which PC1 explains about 80%. Its average score increases from baseline to day 1 (horizontal black arrow in the scores plot of Figure 5B), is approximately stationary at day 2 and decreases, coupled with an increase in PC2, on day 3 (diagonal arrow). The PC1 loading describes a decrease in FS, CD62L and CD16, together with a spreading along the direction of the activation markers (i.e., the horizontally elongated shape of the red density); PC2 is also associated with the spreading of cellular density along several markers. This result is in agreement with the corresponding *top*PCA model examined above: the large patterns of neutrophil activation (PC1 in Figure 4) do not have a correlation with the day-sequence common to the whole studied cohort (though they are likely related to the spreading found in the ASCA loadings), and the *top*PC2 scores also indicated an overall decrease in CD62L and CD16 on days 1–2, followed by a slight increase on day 3. A previous univariate analysis of the same data[37] reached the same conclusions as well. However, the ASCA model delineates the common response to the walking days much more directly and clearly, while avoiding a tedious search throughout the (unsupervised) PCs or the individual markers.

The *day* submodel built on the fMLF+ data explains a larger portion of the data variance (22%, as compared to 12% in the fMLF− model) and is more similar to the full model, both in average score trajectory and *base*PCA loadings, see Figure 5C. Its (ASCA) PC1 loading describes an increase in activation markers (and partially CD16), whereas PC2 is

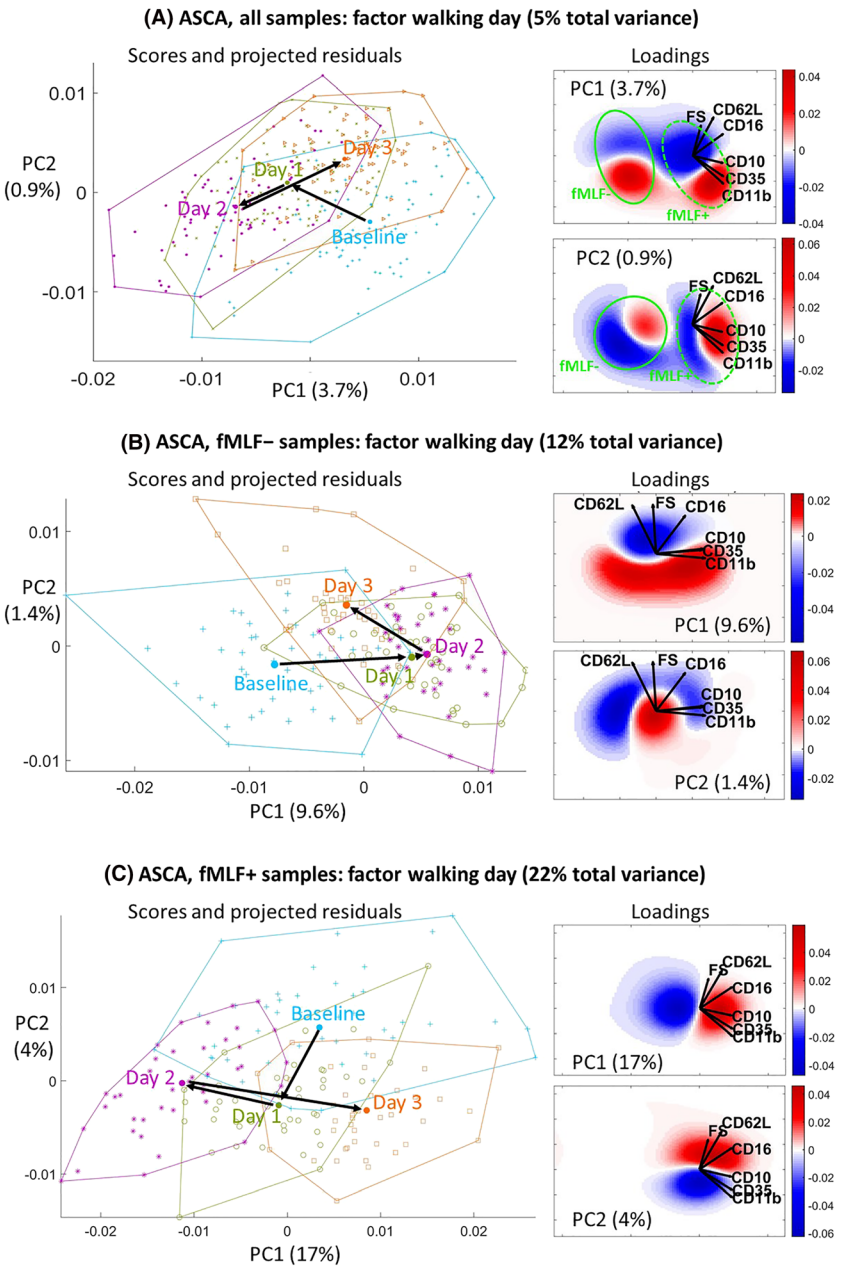

FIGURE 5 ASCA model for factor walking *day*, built on (A) full data set, (B) fMLF− and (C) fMLF+ samples. In the score plots (left side), the larger filled circles correspond to the day averages, whereas the smaller dots are the projected residuals of the model, which are also wrapped by the colored polygons. The black arrows highlight the chronological order of the time points: In all cases the direction towards day 3 is somewhat opposite to the trend from baseline to day 2. In the loadings plot of the full model (A), two shifts in cellular density can be recognized, originating from the fMLF− and fMLF+ samples and encircled with a green continuous and dashed line, respectively. All percentages indicate the proportion of explained variance with respect to the total variance of the data used for the relevant model

aligned towards an increase in CD62L and CD16. This results in a decrease in CD62L and CD16 from baseline to day 1, followed by a decrease in activation markers on day 2 and a subsequent large increase on day 3.

Because ASCA models several groups/levels and factors at once, for a given pair of data subsets it may not be as accurate as methods that model that instance specifically. To get an assessment of how the two outcomes may differ, we compared certain ASCA results with the corresponding ones from DAMACY, see Figure 6. In particular, the plots depict the average shift in cell density between baseline (blue) and day 3 (red), for both fMLF− (top row) and fMLF+ (bottom row). The ASCA plots (lefthand column) are obtained by multiplying the difference in scores (day 3 - baseline)

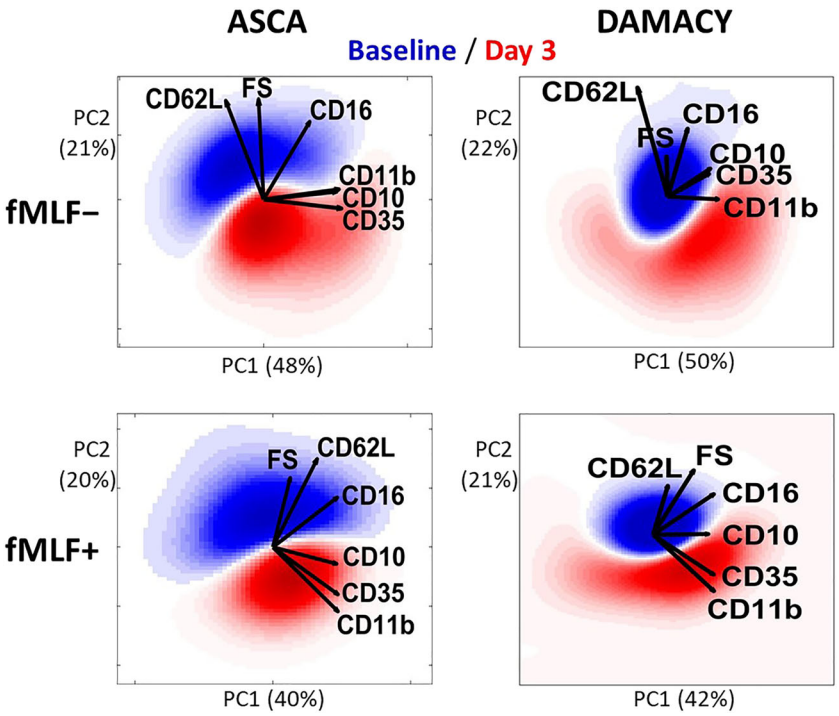

FIGURE 6 Comparison between ASCA (left) and DAMACY (right) results. All plots depict the average shift in cellular density between baseline (blue) and day 3 (red), as calculated for fMLF− (top row) and fMLF+ (bottom row) data. The ASCA plots are obtained by multiplying the difference in scores (day 3 - baseline) by the loadings of the corresponding partial model, using 3 PCs. The DAMACY plots are weight maps. The loading arrows and the percentages of explained variance (in parentheses along the axes) refer to the *base*PCA built on that specific subset. The color scales are analogous to those of the previous figures

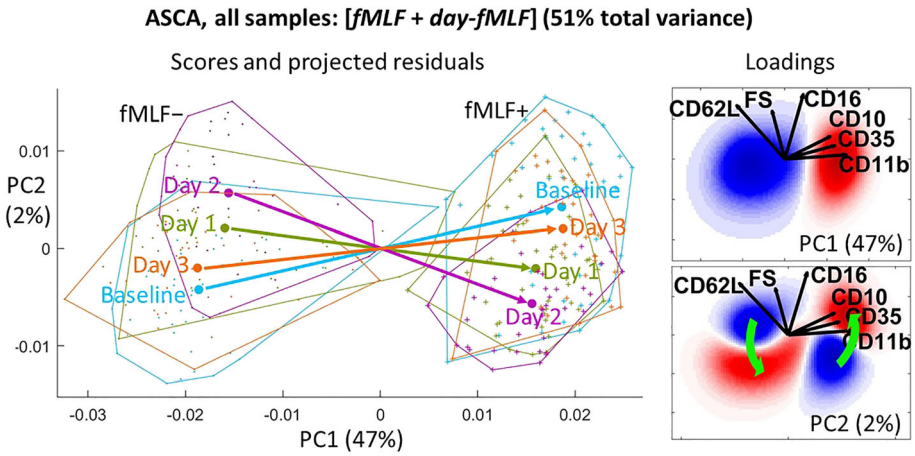

FIGURE 7 ASCA model combining the *fMLF* factor with the *day-fMLF* interaction. Score plots (left-hand side): The large filled circles correspond to the day-fMLF averages, whereas the projected residuals are indicated with small dots (fMLF−) and crosses (fMLF+) and wrapped by the colored polygons. The colored straight arrows represent the average effect of response to the fMLF stimulation for each day. Loading plots (right-hand side): PC1 describes the average fMLF-induced neutrophil activation, whereas PC2 describes a rotation of the direction of cell density shift (green curved arrows). All percentages of explained variance refer to the total data set. The color scales are analogous to those of the previous figures

by the loadings of the corresponding partial model, using three PCs (which is the theoretical maximum). The DAMACY plots (righthand column) are weight maps (i.e., refolded weight vector) from the models that discriminate baseline from day 3 for each fMLF subset; they all had an accuracy >90% and $p < 0.01$. All the loading arrows and the reported percentages of explained variance refer to the *base*PCA models built on the relevant subset.

The plots obtained by ASCA are similar to the ones from DAMACY and the main cell shifts are oriented in the same direction. However, there are some differences in the shape of the cell densities, especially in the peripheral regions originating from samples or cells that behave differently than the majority of the data set. The ASCA model is also affected by the rigidity imposed by a *base*PCA common to all factor levels. This fact is apparent by looking at the DAMACY model for fMLF− (top right), which reveals that in the baseline – day 3 shift the activation markers are not so tightly correlated anymore, and that CD11b presents a higher overall increase than CD10 and CD35 (CD11b points more directly at the red cloud).

3.2.2 | Interaction of *fMLF* with the walking *day*

According to Table 2, the full ASCA model describes two significant interactions with *fMLF*, the largest of which (4.4% of explained variance) is the one with the factor *day*. Because the plots of pure interaction effects are often difficult to interpret, we calculated a submodel combining the *fMLF* factor with the *fMLF-day* interaction as described in Section 2.2.2, see Figure 7. This submodel facilitates the visualization of how the fMLF response changes during the different days. The scores plot shows that on all days most of the fMLF response (colored straight arrows) is explained by PC1, whose loading is associated to a neutrophil activation analogous to the ones seen in Figures 3 and 4, as was expected. On the other hand, the difference in such response among the different days is mainly explained by PC2, which expresses a change in the direction of cell density shift. In particular, an increase in PC2 corresponds to a counter-clockwise rotation of such shift, as pointed out by the green curved arrows. This rotation translates to an activation that is slightly more intense on CD10 and less on CD11b, a condition that occurs most strongly at baseline, decreases on days 1 and 2 and reverts on day 3.

Analogous plots for the interaction between *fMLF* and *statin* (shown in Figure S1) indicate that the fMLF response is slightly less intense for the Statin 2 group. Although the tiny effect size of the *fMLF-statin* interaction (0.4%) warrants caution towards this finding, this result could be the basis for further confirmatory studies, perhaps with a larger cohort.

3.2.3 | Individual stratification of the main walking effect

A drawback of the models presented above is that the ASCA loadings, for example, the ones in Figure 5B, do not indicate whether the plotted average response originates from a single response common to all samples (e.g., all individuals produce both activated and non-activated neutrophils) or instead from a sum of different individual responses (e.g., certain individuals respond with activation, others with non-activation of neutrophils). To shed light on this aspect, it is convenient to calculate the interactions of the *individual* factor with other effects, which were previously left in the residuals of (1). For example, the submodel combining the factor *individual* with the *day-individual* interaction expresses the deviation of each subject from the average walking-day effect (the effect shown in Figure 5). For the fMLF− subset, the first two loadings of this ASCA submodel describe neutrophil activation (PC1) and spreading of cell density along the direction of the activation markers and of CD16 (PC2), see bottom plots of Figure 8. The high variance (88%) explained by this submodel indicates that the individual deviations are quite large as compared to the average response.

To identify the major types of dynamics present in this subset, we performed a cluster analysis of the **T** matrix in a sequential manner as described in Section 2.2.3; the derived dendrograms are plotted in Figure 8. With this sequential approach, the study group is progressively split at each time point, so that, for example, two individuals differing strongly on the first day will fall into two separate clusters even if their responses on subsequent days are very similar. We believe that this clustering criterion is more suited to sort chronological sequences, as opposed to standard clustering algorithms that are not affected by the variable order; other clustering methods may be applied in other situations.

The clustering revealed a stratification of the subjects into seven clusters with distinct response patterns. Their per-day averages are plotted in Figure 8 below the dendrograms, with staggered lines indicating the trajectory from baseline to day 3 according to the direction of the arrow. The color of these trajectories matches that of the rectangle surrounding the corresponding cluster in the last layer of dendrograms (day 3), whereas their thickness is proportional to the number of subjects in the cluster. The trajectories of most individuals, corresponding to the largest clusters plotted on the left, all start and end near the center of the axes, indicating a response that eventually returns to the average state.

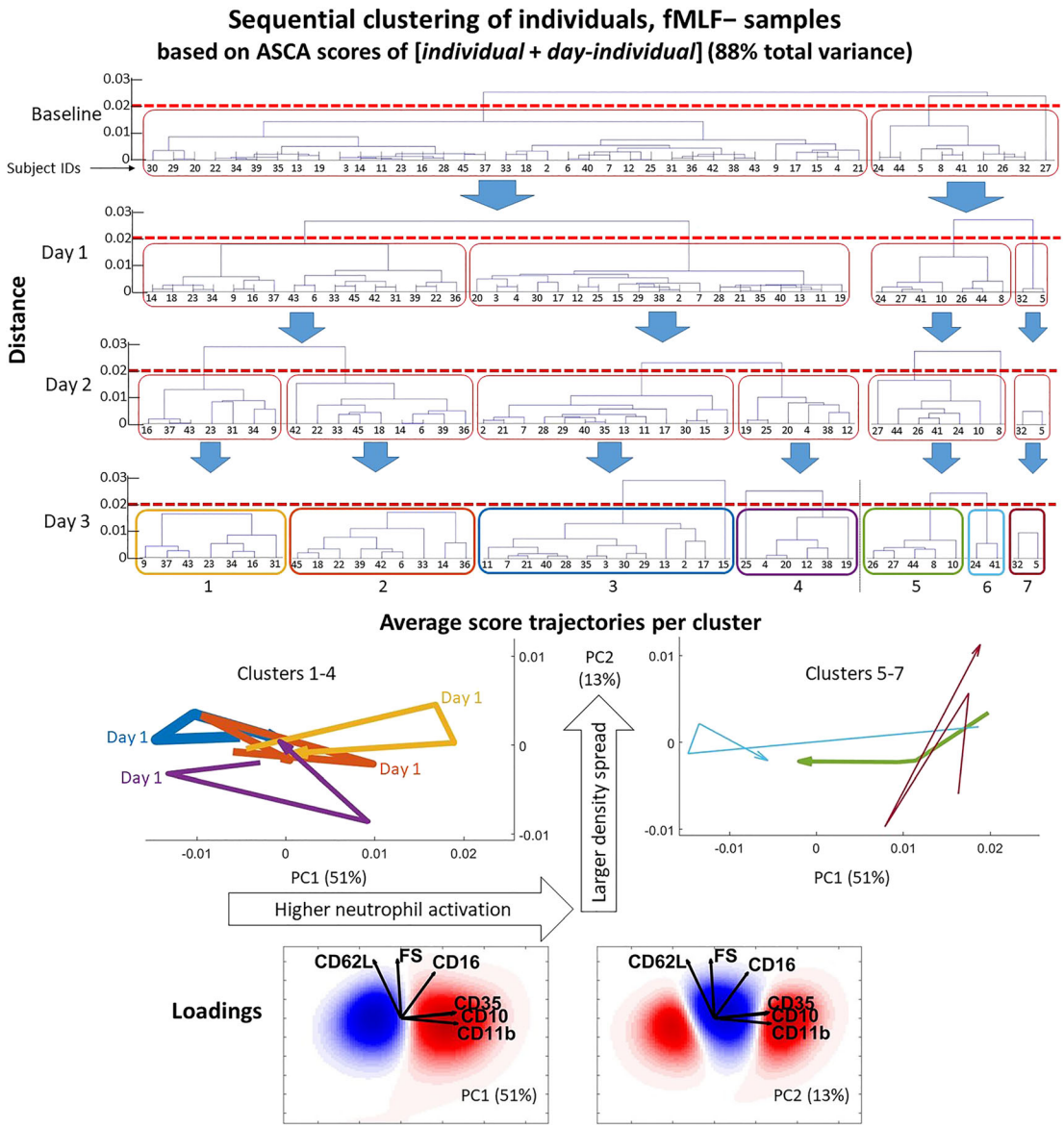

FIGURE 8 Sequential clustering of individuals based on the ASCA scores of the submodel [*individual* + *day-individual*], calculated on the fMLF− data. For each day, new dendrograms are constructed using the scores of PC1 and PC2 relative to that day. The dashed red line indicates the threshold for cluster separation (a few isolated samples are merged with the nearest cluster). Middle plots: Average trajectories of ASCA scores (from baseline to day 3 according to the direction of the arrows) for each final cluster. The color of the lines match those of the rectangles in the bottom layer of dendrograms, and the line thickness is proportional to the cluster size. The clusters 5–7 are shown on a different plot for clarity. Bottom plots: PC1 and PC2 loadings of the considered ASCA submodel, describing neutrophil activation and spreading of cell density, respectively. All percentages of explained variance refer to the fMLF− subset. The color scales are analogous to those of the previous figures

What changes between these clusters is the direction of the response, for example, between baseline and day 1 the yellow and orange clusters proceed towards the right, corresponding to a higher neutrophil activation, whereas the blue and purple clusters go the opposite way. Clusters 5–7, plotted separately on the right, are instead characterized by a high neutrophil activation already at baseline, followed by an evolution to a state considerably different than the initial one.

Despite the trajectories' overlap, it is possible to appreciate the distinct time-patterns, especially with regards to the activation direction (PC1). For example, the yellow and the purple arrows follow opposite paths from baseline to day 1 (segment highlighted by the numbers "0" and "1"), corresponding to higher and lower neutrophil activation, respectively, before returning to a near-initial state on day 3.

This mathematical tool enables a quick recognition of the main patterns of response to walking, highlighting which ones are more common and which are outlying. Such result may facilitate the understanding of which biological processes are prevalent in different subjects, as well as suggest what amount of exercise is most appropriate for each population group. Moreover, this approach, which was here limited to two PCs for ease of visualization, may be extended to further PCs to obtain a more detailed representation of the individual responses. For instance, PC3 (see Figure S2) describes a deformation of the cellular density along a bean-like shape, which is a well-known type of distribution already observed in neutrophils.[17]

Finally, a further improvement on this methodology, particularly in assessing differences among individuals, could be to employ multivariate mixed modeling techniques that are able to determine random effects. Two recently published methods, Linear Mixed Model-PCA (LiMM-PCA)[52] and Repeated Measures-ASCA+ (RM-ASCA+),[53] appear very promising candidates for development in this direction.

4 | CONCLUSION

Multicolor Flow Cytometry is nowadays a mature and extremely valuable technique, increasingly widespread in the medical and biological field. Nevertheless, the size and complexity of the data that it produces, even with a low number of markers, still poses challenges in the detection, visualization and interpretation of the relevant distributions patterns of cell phenotypes. Ultimately, the full exploitation of the wealth of information that can be obtained from MFC depends on the availability of appropriate mathematical tools.

In this work, we proposed a further option for the analysis of MFC data, consisting in an implementation of ASCA based on previous linear multivariate MFC-dedicated methods, which have the advantage of showing explicitly the direction of biological variability of the highlighted cells. Compared to these methods, ASCA provides a much more comprehensive insight in the immunological variation associated with different factors and interactions in the experimental design, enabling transparency from changes on individual cells to systematic response-associated immunological changes on given time points. In our case study on the innate immune response to prolonged physical exercise, ASCA was able to reveal the effects on the cell populations caused by the known experimental factors and their interactions, directly quantifying their size and statistical significance. Most relevant were the average effect of exercise (an initial neutrophil mobilization on the first 2 days of walking, followed by a partial reversal on the third day) and the relationship between exercise and response to bacterial stimulation (fMLF). Using an appropriate time-guided clustering method, we also found a stratification of individual dynamics, for example, some individuals had a high neutrophil activation on day 1 and 2 before returning to the initial state on day 3, whereas another group of individuals had the opposite dynamic. Some of these findings were related to small effects that are normally hidden by larger sources of variation and are thus difficult to detect using only PCA, which does not take the information on the experimental design into account. This work also compared ASCA results with those from a discriminant analysis method (DAMACY), showing that the former may be less accurate at describing differences between specific groups of data, but it is more efficient at providing a general and simultaneous description of all the groups involved in the experimental design. This characteristic is especially useful to analyze sequences (such as a time series in this case), as opposed to pairwise classifications.

The results of this work show that the proposed implementation of ASCA is an effective way to disentangle variation in MFC measurements, particularly changes in relative expression of several cell markers simultaneously, caused by multifactorial experimental designs. It therefore constitutes a valuable addition to the mix of chemometric techniques to be employed for a comprehensive multivariate analysis of this type of data.

ACKNOWLEDGEMENTS

This work is part of the research programme 'NWA startimpuls meten en detecteren van gezond gedrag' with project number 400.17.604, which is (partly) financed by the Dutch Research Council (NWO).

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/cem.3402.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Carlo G. Bertinetto https://orcid.org/0000-0003-1728-396X

REFERENCES

1. Brown M, Wittwer C. Flow cytometry: principles and clinical applications in hematology. *Clin Chem*. 2000;46(8):1221-1229. doi:10.1093/clinchem/46.8.1221
2. Robinson JP, Roederer M. Flow cytometry strikes gold. *Science*. 2015;350(6262):739-740. doi:10.1126/science.aad6770
3. Laerum OD, Farsund T. Clinical application of flow cytometry: a review. *Cytometry*. 1981;2(1):1-13. doi:10.1002/cyto.990020102
4. Spijkerman R, Hesselink L, Hellebrekers P, et al. Automated flow cytometry enables high performance point-of-care analysis of leukocyte phenotypes. *J Immunol Methods*. 2019;474:112646. doi:10.1016/j.jim.2019.112646
5. Lugli E, Roederer M, Cossarizza A. Data analysis in flow cytometry: the future just started. *Cytom Part a*. 2010;77A(7):705-713. doi:10.1002/cyto.a.20901
6. Amir EAD, Davis KL, Tadmor MD, et al. ViSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol*. 2013;31(6):545-552. doi:10.1038/nbt.2594
7. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008:2579-2605.
8. Van Gassen S, Callebaut B, Van Helden MJ, et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytom Part A*. 2015;87(7):636-645. doi:10.1002/cyto.a.22625
9. Qiu P, Simonds EF, Bendall SC, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol*. 2011;29(10):886-891. doi:10.1038/nbt.1991
10. Li Y, Mahjoubfar A, Chen CL, Niazi KR, Pei L, Jalali B. Deep cytometry: deep learning with real-time inference in cell sorting and flow cytometry. *Sci rep*. 2019;9(1):11088. doi:10.1038/s41598-019-47193-6
11. Tinnevelt GH, Jansen JJ. Resolving complex hierarchies in chemical mixtures: how chemometrics may serve in understanding the immune system. *Faraday Discuss*. 2019;218(0):317-338. doi:10.1039/c9fd00004f
12. Tinnevelt GH, Kokla M, Hilvering B, 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):1-11. doi:10.1038/s41598-017-05714-1
13. Jansen JJ, Hilvering B, van den Doel A, et al. FLOOD: FLow cytometric Orthogonal Orientation for Diagnosis. *Chemom Intel Lab Syst*. 2016;151:126-135. doi:10.1016/j.chemolab.2015.12.001
14. Folcarelli R, Van Staveren S, Bouman R, et al. Automated flow cytometric identification of disease-specific cells by the ECLIPSE algorithm. *Sci rep*. 2018;8(1):1-18. doi:10.1038/s41598-018-29367-w
15. Folcarelli R, Tinnevelt GH, Hilvering B, et al. Multi-set pre-processing of multicolor flow cytometry data. *Sci rep*. 2020;10(1):1-12. doi:10.1038/s41598-020-66195-3
16. Tinnevelt GH, Van Staveren S, Wouters K, et al. A novel data fusion method for the effective analysis of multiple panels of flow cytometry data. 2019:1-9. doi:10.1038/s41598-019-43166-x
17. Van Staveren S, Ten Haaf T, Klöpping M, et al. Multi-dimensional flow cytometry analysis reveals increasing changes in the systemic neutrophil compartment during seven consecutive days of endurance exercise. *PLoS One*. 2018;13(10):1-23. doi:10.1371/journal.pone.0206175
18. Jansen JJ, Hoefsloot HCJ, van der Greef J, Timmerman ME, Westerhuis JA, Smilde AK. ASCA: Analysis of multivariate data obtained from an experimental design. *J Chemometr*. 2005;19(9):469-481. doi:10.1002/cem.952
19. Bertinetto C, Engel J, Jansen J. ANOVA simultaneous component analysis: a tutorial review. *Anal Chim Acta X*. 2020;6:100061. doi:10.1016/j.acax.2020.100061
20. Grassi S, Lyndgaard CB, Rasmussen MA, Amigo JM. Interval ANOVA simultaneous component analysis (i-ASCA) applied to spectroscopic data to study the effect of fundamental fermentation variables in beer fermentation metabolites. *Chemom Intel Lab Syst*. 2017;163:86-93. doi:10.1016/j.chemolab.2017.02.010
21. Liland KH, Smilde A, Marini F, Næs T. Confidence ellipsoids for ASCA models based on multivariate regression theory. *J Chemometr*. 2018;32(5):1-13. doi:10.1002/cem.2990
22. Ryckewaert M, Gorretta N, Henriot F, Marini F, Roger JM. Reduction of repeatability error for analysis of variance-simultaneous component analysis (REP-ASCA): application to NIR spectroscopy on coffee sample. *Anal Chim Acta*. 2020;1101:23-31. doi:10.1016/j.aca.2019.12.024
23. Nueda MJ, Conesa A, Westerhuis JA, et al. Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA. *Bioinformatics*. 2007;23(14):1792-1800. doi:10.1093/bioinformatics/btm251
24. Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers RJAN, van der Greef J, Timmerman ME. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*. 2005;21(13):3043-3048. doi:10.1093/bioinformatics/bti476
25. Chang W-T, Thissen U, Ehlert KA, et al. Effects of growth conditions and processing on rehmannia glutinosa using fingerprint strategy. *Planta Med*. 2006;72(5):458-467. doi:10.1055/s-2005-916241
26. Saccenti E, Smilde AK, Camacho J. Group-wise ANOVA simultaneous component analysis for designed omics experiments. *Metabolomics*. 2018;14(6):1-18. doi:10.1007/s11306-018-1369-1
27. Babamoradi H, Amigo JM, Van den Berg F, Petersen MR, Satake N, Boe-hansen G. Quality assessment of boar semen by multivariate analysis of fl ow cytometric data. *Chemom Intel Lab Syst*. 2015;142:219-230. doi:10.1016/j.chemolab.2015.02.008

28. Spijkerman R, Hesselink L, Bertinetto C, et al. Analysis of human neutrophil phenotypes as biomarker to monitor exercise-induced immune changes. *J Leukoc Biol*. 2020;109(4):833-842. doi:10.1002/JLB.5A0820-436R
29. Shephard RJ. Development of the discipline of exercise immunology. *Exerc Immunol Rev*. 2010;16:194-222.
30. Nieman DC, Wentz LM. The compelling link between physical activity and the bodys defense system. *J Sport Heal Sci*. 2019;8(3):201-217. doi:10.1016/j.jshs.2018.09.009
31. van der Geest KSM, Wang Q, Eijsvogels TMH, et al. Changes in peripheral immune cell numbers and functions in octogenarian walkers - an acute exercise study. *Immun Ageing*. 2017;14(1):1-13. doi:10.1186/s12979-017-0087-2
32. Spijkerman R, Hesselink L, Bertinetto C, et al. Refractory neutrophils and monocytes in patients with inflammatory bowel disease after repeated bouts of prolonged exercise. *Cytometry*. 2021;100:676-682. doi:10.1002/cyto.b.21996
33. Nieman DC. Exercise, infection, and immunity. *Int J Sports Med*. 1994;15(S 3):S131-S141. doi:10.1055/s-2007-1021128
34. Schwellnus M, Soligard T, Alonso JM, et al. How much is too much? (Part 2) International Olympic Committee Consensus Statement on load in sport and risk of illness. *Br J Sports Med*. 2016(17):1043-1052. doi:10.1136/bjsports-2016-096572
35. Gleeson M, Bishop N, Walsh N. Exercise immunology. 2013. doi:10.4324/9780203126417
36. Alenghat FJ, Davis AM. Management of blood cholesterol. *JAMA - J Am Med Assoc*. 2019;321(8):800-801. doi:10.1001/jama.2019.0015
37. Spijkerman R, Hesselink L, Bertinetto C, et al. Analysis of human neutrophil phenotypes as biomarker to monitor exercise-induced immune changes. *J Leukoc Biol*. 2021;109(4):833-842. doi:10.1002/JLB.5A0820-436R
38. Panaro MA, Mitolo V. Cellular responses to FMLP challenging: a mini-review. *Immunopharmacol Immunotoxicol*. 1999;21(3):397-419. doi:10.3109/08923979909007117
39. Pillay J, Kamp VM, Van Hoffen E, et al. A subset of neutrophils in human systemic inflammation inhibits T cell responses through Mac-1. *J Clin Invest*. 2012;122(1):327-336. doi:10.1172/JCI57990
40. Folcarelli R, Van Staveren S. Transformation of multicolour flow cytometry data with OTflow prevents misleading multivariate analysis results and incorrect immunological conclusions. doi:10.1002/cyto.a.24491
41. Van Deun K, Smilde AK, van der Werf MJ, Kiers HAL, Van Mechelen I. A structured overview of simultaneous component based data integration. *BMC Bioinformatics*. 2009;10(1):246. doi:10.1186/1471-2105-10-246
42. Kiers HAL, ten Berge JMF. Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure. *Br J Math Stat Psychol*. 1994;47(1):109-126. doi:10.1111/j.2044-8317.1994.tb01027.x
43. Ten Berge, JMF, Kiers HAL, Van der Stel V. Simultaneous components analysis. *Stat Appl*. 1992;4(4):377-392.
44. Marini F, de Beer D, Joubert E, Walczak B. Analysis of variance of designed chromatographic data sets: the analysis of variance-target projection approach. *J Chromatogr A*. 2015;1405:94-102. doi:10.1016/j.chroma.2015.05.060
45. Westerhuis JA, van Velzen EJJ, Hoefsloot HCJ, Smilde AK. Multivariate paired data analysis: Multilevel PLSDA versus OPLSDA. *Metabolomics*. 2010;6(1):119-128. doi:10.1007/s11306-009-0185-z
46. Thiel M, Féraud B, Govaerts B. ASCA+ and APCA+: extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. *J Chemometr*. 2017;31(6):1-13. doi:10.1002/cem.2895
47. Vis DJ, Westerhuis JA, Smilde AK, van der Greef J. Statistical validation of megavariate effects in ASCA. *BMC Bioinformatics*. 2007;8(1):1-8. doi:10.1186/1471-2105-8-322
48. Bylesjö M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J Chemometr*. 2006;20(8-10):341-351. doi:10.1002/cem.1006
49. Folch-Fortuny A, Arteaga F, Ferrer A. Missing data imputation toolbox for MATLAB. *Chemom Intel Lab Syst*. 2016;154:93-100. doi:10.1016/j.chemolab.2016.03.019
50. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. In: *ACM Computing Surveys*. Vol. 31; 1999:264-323. doi:10.1145/331499.331504.
51. Podani J, Miklós I. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology*. 2002;83(12):3331-3343. doi:10.1890/0012-9658(2002)083[3331:RCATHE]2.0.CO;2
52. Martin M, Govaerts B. LiMM-PCA: Combining ASCA+ and linear mixed models to analyse high-dimensional designed data. *J Chemometr*. 2019;2020(6):1-20. doi:10.1002/cem.3232
53. Madssen TS, Giskeødegård GF, Smilde AK, Westerhuis JA. Repeated measures ASCA+ for analysis of longitudinal intervention studies with multivariate outcome data. *PLoS Comput Biol*. 2021;17(11):1-21. doi:10.1371/journal.pcbi.1009585

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: Bertinetto CG, Spijkerman R, Hesselink L, et al. Comprehensive multivariate evaluation of the effects on cell phenotypes in multicolor flow cytometry data using ANOVA simultaneous component analysis. *Journal of Chemometrics*. 2023;37(7):e3402. doi:10.1002/cem.3402