Qeios

Peer Review

Review of: "CAFE: An Integrated Web App for High-Dimensional Analysis and Visualisation in Spectral Flow Cytometry"

Christophe Pellefigues[1]

1. French Institute of Health and Medical Research, Paris, France

Siam et al. present an interesting new open-access app designed to facilitate the analysis of high-dimensional spectral flow cytometry (SFCM) data. The flow cytometry community definitely needs more good open-access apps to analyze SFCM, especially locally without an internet connection. This new tool seems promising, easy to use, and powerful. However, CAFE needs to be described in more detail, to have some quality control, and a more thorough validation, especially in comparison with previously published results.

Major points:

1. In order to validate the functionality of the app, the authors should compare their analysis to published results of the dataset they are importing. I sincerely believe that a proper validation requires analyzing at least two independent datasets and showing that CAFE is able to provide similar or even better results than the main published results. The validation does not need to be extensive, but the authors need to compare their results to previously obtained results.

For the current dataset, the authors should show that CAFE is able to provide the same results as the ones presented in Yu et al., such as a loss of CD8+ CD161+ T cells with COVID severity, and the other results taken from this particular dataset. Some cell subsets have not been detected after CAFE UMAP annotation, for example, (i.e., PMN). Some other subsets seem underrepresented in CAFE (i.e., basophils). Can CAFE identify rare subsets such as B cell or NK subsets, such as in Yu et al.? Why are some particular subsets underrepresented?

 doi.org/10.32388/3M1I3X

Using the same terminology is paramount (i.e., MAIT vs CD8+ CD161+). What is the percentage of CD8+ TCRab+ among $CD161^{hi}$ cells in each group, and how does that compare to the initial analysis (Figure 1G, Yu et al.)? Can the authors find the same differences between healthy males and females in CD8+ (Figure 2, Yu et al.)? Discriminating the UMAP results obtained in CAFE based on patient groups (healthy, exposed, infected, hospitalized), and highlighting clearly the CD8+ CD161+ population on the UMAP, is essential (i.e., here it is difficult to ascertain which is the MAIT cluster or the naïve CD8, etc.). Figure 3 should be changed accordingly. Additionally, plotting cell counts obtained from PBMCs is not relevant due to strong changes in interindividual recovery of the Ficoll ring (Figure 3c, 4e, 4f). Percentages of PBMC should be used instead. In Figure 3f, proportions of cells are described in the legend, but counts are presented.

The n of data is inconsistent between the various presented results. Yu et al. show n=9-10 for healthy vs. hospitalized. The authors state that they used 10 samples with n=5 per group. Yet Figure 3f shows n=9 per group. Counting the dots in 3a or 3f, it seems that this is closer to n=7? Only Figures 4 d, e seem limited to n=5. Either way, it would be impossible to validate that CAFE can reproduce the original analysis if the authors do not analyze all the individuals from the original analysis. It is likely that the authors would get even better results by analyzing all the individuals from all the groups from the initial analysis.

1. Most FCM intensity data are usually distributed in a logarithmic fashion and need to be analyzed as such. The bi-exponential display is actually the best way to represent FCM intensities and to isolate clusters based on marker expression (arcsinh transformation). For example, noise values may range from 10^{-4} to 10^{4} for a particular marker, with specific values only being detected above 10^{5}. This is why proper data scaling is critical before feeding the dataset to any algorithm. In the above example, depending on how the data is scaled and normalized, the 10^{4} intensities may be considered null (they should) or extremely positive (they should not), and 10^{5} should clearly be considered as a positive intensity above a negative cluster ranging from 10^{-4} to 10^{4}. Any other analysis would actually generate analysis-driven artefacts.

While I perfectly understand that CAFE was not meant for SFCM data preprocessing and scaling, I underline this point here because the dataset chosen to validate CAFE, or the way the data of the dataset was scaled, is impeding CAFE from doing a good analysis.

Here, CD19 values seem poorly scaled, for example, and there's no telling that the B cell cluster shows CD19 positivity on UMAP (Figure 2A). In properly designed experiments, all negative intensity values should arise from the uncertainty surrounding noise detection. This means that they should follow a relatively normal distribution around a median null to positive value. Here, the CD19 data strains a lot in the negative values range, which could be linked to several common causes: 1) Unmixing issues that overcompensate CD19; 2) Outliers with an extremely negative value were kept in the analysis and completely messed up the data scaling (maybe in the cDC cluster). Fixing the cause is important to improve any subsequent dimensionality reduction analysis and data interpretation.

Moreover, for all the UMAP representations, changing the scaling of the data, or the visual scale of intensities, would do wonders to better interpret the data and visualize positive from negative populations (i.e., CD3, CD11b, HLADR, CD20, and IgM). It seems here the color scale was applied automatically and was not manually adjusted, which can lead to misinterpretation if the data was poorly scaled in the first place. It is also important for all the marker expressions used for the UMAP analysis to be represented in the manuscript, but only 15 are represented here. Importantly, some markers can be expressed on all cells of a dataset at a relatively high level, with only minimal differences between populations. In this dataset, this is the case for CD45, as the dataset has been preprocessed as CD45+ PBMCs viable singlets. If this marker is not scaled properly, it will only add some more noise to the dimensionality reduction analysis. If analyzing CD45 may be interesting in some particular experiments containing CD45 low or negative cells, it is often excluded from dimensionality reduction analyses in datasets containing only CD45+ cells.

Scaling errors or artefacts can lead to wrong assumptions down the pipeline and to the wrong identification of cell clusters, for example. I recommend properly scaling all the SFCM data (and maybe removing CD45) before running CAFE. The authors also need to state exactly which parameters have been selected to run the analysis and which were discarded (i.e., FSC-A, FSC-H…), and why.

Minor points:

1. Loss of rare events can be a major drawback in discovering new cell subsets. Filtering out data should ideally be limited to preprocessing and to strictly removing artefacts. By nature, a lot of interesting signals may be outliers in the extremum of a Gaussian distribution. So I have a lot of reservations concerning any automated "noise" reduction protocols. Can you provide an unambiguous QC of ComBat applied to SFC data and a QC of PCA-KDE based downsampling when run in CAFE? What was removed from the data? What happened during batch homogenization?

2. The dendrogram representation of the mean expression of the various markers seems to be scaled differently than the UMAP representation. For example, CD19 expression seems renormalized but based on the same data with a scaling problem, and the dendrogram determined that most cells of the dataset are CD19+, which is evidently false from a biological point of view. Can you explain how the intensities were transformed between the two analyses? It seems that the positivity was determined solely based on the flawed data distribution. This can lead to serious misinterpretation of the data.
3. The automatic identification of the basophil subset seems absolutely flawed. Basophils do not express CD3 nor CD4, while here they show a very high expression of both in both UMAP and dendrogram. They are not known to express T cell levels of CCR7 or CD27 at steady state in human blood neither. The only proper fitting phenotype is that they express CD123 but not HLADR. Similarly, the pDC cell cluster seems flawed as well: pDCs are usually identified as $CD123^{hi}$ $HLADR^{hi}$, but most of them here seem to lack HLADR expression. For both of these populations, it is difficult to assess if either the data scaling of marker intensities is flawed or if the cluster identification was wrong. The authors should determine the cause of this issue (which may again be only a problem with data scaling in the first place).
4. I do not understand how easily a user can export data from clusters in a common format like xls or csv. Many users will want to save intensity geometric means or medians or the frequency of a particular subset from various subsets, for example, to use with data from other experiments in third-party statistics software. What kind of data can be exported by CAFE? Figure 1 shows counts and frequencies, but this is rather limited. However, the authors explain that "CAFE outputs a file with expression data for all markers on each cluster by sample." What exactly can be exported should be described.
5. CAFE can perform various statistical tests depending on the normality of the distribution of the data, which is smart. However, in the current manuscript, the statistical tests used are not described. Figure 4a compares the expression of CD161 in different cell types. The authors need to describe which cells they are using. Cells from only healthy persons? Has the data been downsampled to a constant number of each subset per patient prior to pooling the data? Does each dot correspond to an individual cell or to the median or the mean of the subset from each individual? Why are there some extreme negative values? Additionally, in order to affirm that MAIT cells show the highest expression of CD161, the authors need to perform a statistical test to

compare the expression between the cell subsets. Figure 4c displays a "median expression of markers across all cell types between COVID19 and healthy individuals, where positive values indicate upregulation." I fail to understand what the reference point is here. Usually, it would have been the healthy group, but they are plotted as well and different from 0 or 1. How was the normalization done? What is exactly represented here? Either way, individual values and a SD or SEM or CI should be represented, and statistical tests comparing the healthy to the COVID19 group, for each marker, should be conducted.

6. As the total number of B or CD8+ T cells recovered from each patient can be very heterogeneous in PBMCs, and especially when comparing with COVID19 patients, it is very difficult to interpret B cell counts as positive or not for a particular marker. It would be better if these values were normalized on the total number of B cells for each individual.

Declarations

Potential competing interests: No potential competing interests to declare.