Cytofast: A workflow for visual and quantitative analysis of flow and mass cytometry data to discover immune signatures and correlations

Guillaume Beyrend a,⁎, Koen Stam b, Thomas Höltt c, Ferry Ossendorp a, Ramon Arens a

a Department of Immunohematology and Blood Transfusion, Leiden University Medical Center, Albinusdreef 2, Leiden 2333 ZA, The Netherlands
b Department of Parasitology, Leiden University Medical Center, Leiden, The Netherlands
c Department of Computer Graphics and Visualization Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, the Netherlands

Article history:
Received 12 July 2018
Received in revised form 6 October 2018
Accepted 12 October 2018
Available online xxx

Keywords:
Flow and mass cytometry analysis
R-based package
High-dimensional analysis
Single-cell analysis
Visualization tool

1. Introduction

Mass cytometry (cytometry by time-of-flight; CyTOF) can detect as many as 40 markers present on millions of single cells. The number of studies based on mass cytometry has considerably increased over the last few years, broadening simultaneously the choice of clustering techniques, such as SPADE [1], FlowMaps [2], FlowSOM [3], Phenograph [4], VortexX [5] and Scaffold maps [6], but also dimensionality reduction-based techniques including PCA [7], t-SNE [8] and Diffusion Map [9]. A recent new computational tool, HSNE [10], embedded in Cytoplore [11], proposed a combination of the two afore-mentioned techniques, building a hierarchical representation of the complete data that preserves the non-linear high-dimensional relationships between cells and avoids any down-sampling [12]. The aforementioned computational tools lack, however, the means for downstream analysis, where no automatic process has been proposed to link clusters abundance with e.g. clinical outcome and to visually interpret the data in the context of the experimental setup. Citrus [13] is currently a broadly used analysis tool for statistical comparison but presents several limitations. The number of samples to be included in such analysis should be more than eight and the number of cells are randomly downsampled before analysis. Moreover, Citrus is using agglomerative clustering, which brings complexity in data visualization.

Here, we developed a workflow, called cytofast, for a fast and quantitative analysis of flow and mass cytometry data. Cytofast allows the visualization of cluster phenotypes, their abundance per sample and per group and additionally enables statistical comparisons taking in account different clinical outcome variables.

We verified our workflow on a non-paired mass cytometry dataset originating from a published study focused on differences between effective and ineffective treatment. We also demonstrated the use of cytofast on a paired dataset of a prime-boost vaccination study. In addition, we verified cytofast on a flow cytometry dataset. Together these analyses showed that our workflow is valid, replicating similar findings previously described and in addition provided a deeper exploration of the data by newly identifying cell clusters that correlate to treatment.

2. Results and discussion

2.1. Cytofast: workflow presentation

We designed an R package, named cytofast, and introduced a workflow to quantify and identify significant group-related 83
subsets. Cytofast can be used after cluster analysis, for example with Cytoplore or FlowSOM, has been performed. Here, we focused on the clustering analysis with Cytoplore using mass and flow cytometry datasets.

The workflow of cytofast can be divided in four parts (Fig. 1). First, a heatmap with a dendrogram is generated showing the median ArcSinh-transformed marker expression values (blue-to-red scale) for all the identified clusters (cluster phenotype overview). Second, a quantitative analysis of the clusters is performed, including the relative abundance of each cluster per sample and group. Third, a sample t-SNE analysis is done to visualize the distribution of cells across clusters. Finally, statistical comparisons are made between groups to identify significant differences.
Fig. 2. Identification and abundance of CD45+ cell clusters by cytofast. (A) Mass cytometry data of CD45+ immune cells in the spleen of mice that received effective and ineffective treatment at two different timepoints (Day 3 and Day 8). Heatmap of all 31 CD45+ cell clusters identified independent of treatment based on Cytoplore clustering. Level of ArcSinh-transformed expression marker is displayed by a blue-to-red scale. Dendrogram on the top represents the hierarchical similarity between the identified clusters and is based on Euclidean distance and complete linkage clustering. (B) Heatmap of relative abundance (expressed as variance or dispersion from the mean) for each cluster identified above in each individual mouse. One row is representing one mouse sample, subsets are displayed per column. A green or a purple square is representative of a lower or a higher number of cells, respectively, compared to the average. The dendrogram displayed on the left is based on hierarchical clustering using Euclidean distance and complete linkage clustering. (C) Summary of the data represented on the panel above by using the median for each group.
Fig. 3. Quantitative comparison of cell clusters from unpaired samples by cytofast. (A) Sample t-SNE of the data of Fig. 2 where one dot is representing one sample, based on the cluster frequencies. The two groups, ineffective and effective, can be clearly distinguished, showing that the immune system is differently shaped by treatment efficiency. (B) Bar graphs representing the average of each sample per group overlaid with the dot plots representing the percentage of the cluster for each sample. P-values are provided to indicate significant differences between the ineffective and the effective group for both timepoints. Statistical analysis was performed on individual clusters (annotated from 1 to 31) using a t-test.
heat map is generated showing the cell frequency calculated for each cluster stratified per individual sample. Thus, one row is representing one biological sample and the identified subsets are displayed per column (cluster abundance per sample). A dendrogram on the top represents the hierarchical similarity between the identified clusters. Dendrogram displayed above is based on hierarchical clustering using Euclidean distance and complete linkage clustering. (B) Heatmap of relative abundance (expressed as variance or dispersion from the mean) for each cluster identified above in each individual macaque. One row is representing one macaque blood sample, subsets are displayed per column. A green or a purple square is representative of respectively a lower or a higher number of cells compared to the average. Dendrogram displayed on the left is based on hierarchical clustering using Euclidean distance and complete linkage clustering. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Fig. 5. Quantitative comparison of cell clusters from paired samples by cytofast. (A) Sample t-SNE where one dot is representing one sample, based on the sample frequencies. The two groups can be clearly distinguished, showing that the immune system is differently shaped upon the two immunizations. (B) Box plots representing the average and standard deviation of each sample per group. Samples are paired. P-values are provided to indicate the significance between different groups. Statistical analysis was performed for each individual cluster (annotated from 1 to 23) using a t-test.
2.4. Cytofast applied to paired mass cytometry datasets: effect of prime-boost vaccination

To analyze paired samples by cytofast, we selected a partial dataset from Palgen et al. [16]. The mass cytometry data consists of blood immune cells analyzed one day after first and second immunization of cynomolgus macaques with modified vaccinia virus Ankara (MVA).

The clustering analysis from Cytosplore identified twenty-three clusters, whose phenotype were presented by the heatmap (Fig. 4A). The abundance of each cluster stratified per sample showed a clear distinction between the immune response after prime and boost (Fig. 4B).

The difference between the two groups are clearly represented by the sample t-SNE map (Fig. 5A). The two populations after prime and boost are distinct, and do not cluster per individual, revealing that the variation triggered on the immune system of each individual is stronger than their intrinsic immune signature. The average abundance of each cluster per group can be displayed in bar graphs, and such quantification of immune subsets that relate to therapy.

2.4. Cytofast applied to flow cytometry data.

With proper adjustment of the clustering in Cytosplore, cytofast can also be applied to flow cytometry datasets. We tested here a flow cytometry dataset comparing phenotypes of blood cells between patients suffering from acute myeloid leukemia (AML) and non-affected patients [17] (downloaded from FlowRepository [18]). After processing six random samples from each group (non-affected patients called Normal, and affected patients called AML), we applied cytofast and obtained a similar visualization output to mass cytometry data (Fig. S1). Thus, cytofast is also operational for visualization of complex flow cytometry data sets and can rapidly identify immune subset changes between groups.

3. Conclusion

Here we report the helpfulness of an R-based workflow, named cytofast, which is designed for visual and quantitative analysis of flow and mass cytometry data to discover immune signatures and correlations. We used a non-paired dataset generated by Spitzer and colleagues in which the effects of immunotherapy were examined, and newly identified with cytofast two macrophage clusters (F4/80+/Ly6G-/CD44+/IgD- and F4/80+/MHC-II+/Siglec-F+) that were significantly reduced upon effective therapy, specifically at a later stage. In a prophylactic vaccine study by Palgen et al., paired samples were analysed and showed clear booster effects of the vaccine on myeloid cell subsets. The usefulness of cytofast to distinguish different immune signatures between groups was also observed with flow cytometry datasets of immune cells in the blood of leukemia patients and non-affected individuals.

Some of the displayed clusters appear highly similar in the heatmaps. It is optional for the user to merge similar clusters. Such clustering merging might remove subtleties, yet this might also make the results more robust against individual sample variation or batch effects.

For all tested datasets, cytofast yielded rapid results. With standard modern computer hardware and specifications (32GB RAM memory), the output was produced in less than thirty seconds (excluding clustering analysis with e.g. Cytosplore), making this tool highly suitable for a rapid and visual data screening.

4. Material and methods

4.1. Use of clustering method

The two cohort mass cytometry datasets were downloaded from Cytobank.

4.1.1. Use of cytofast

The files were uploaded to Cytosplore, already pre-gated by the original authors. Files were sample-tagged by adding the CSPRL_ST channel, their marker expression arcsinh5 transformed and subjected to dimensionality reduction. The iterations chosen of the HSNE analysis were 1,000. We clustered the data with a kernel size sigma of 30 on the overview level and exported the resulting clusters without manually modifying it. Lastly, in R we used cytofast for further analysis of those files.

4.1.2. Use of other clustering method

The use of cytofast is not restricted to a pre-processing of the data with Cytosplore. Most of the clustering providing a cluster assignment to each cell can be used (e.g. FlowSOM as detailed in the vignette).

4.2. Use of cytofast

Our workflow is linked to the use of an upstream clustering tools such as Cytospore or FlowSOM. We load the clusters produced by applying the HSNE dimensionality reductions and mean shift clustering in Cytospore and saved as FCS files. After loading the FCS files of all clusters, cytofast analysis is based on the sample characteristics written to a single file, gathering relevant information like sample name, clinical outcome and sample tag (also named CSPRL_ST in Cytospore). The
clinical input can be any qualitative clinical data like gender, or group affiliation, or quantitative like tumor size or age.

The vignette, where all the steps are explained to facilitate the reproducibility of the figures and the package itself will be deposited on Bioconductor. A downsampled number of cells from the Spitzer study is included in the vignette. Differences might be seen between the downsampled dataset provided in the vignette and the analysis from this paper due to the downsampling.

4.3. Flow cytometry data

We confirmed that our R script was also valid to represent flow cytometry datasets. We downloaded data files from FlowRepository (https://flowrepository.org/id/FR-FCM-ZZYA) and selected six patients per group (Files 0006.FCS, 0014.FCS, 0022.FCS, 0030.FCS, 0046.FCS, 0062.FCS from the Normal group; Files 0038.FCS, 0070.FCS, 0206.FCS, 0262.FCS, 0294.FCS, 0390.FCS from the AML group). Data was clustered with Cytosplore, clusters sharing high similarities were manually merged and the resulted output were processed with cytofast.

Author contributions

GB wrote the manuscript and the workflow. KS wrote and embedded the workflow into the R package cytofast. TH reviewed the manuscript and gave conceptual input. FO gave conceptual input. RA wrote the manuscript and supervised the project.

Conflict of interest

The authors declare no financial or commercial conflict of interest.

Acknowledgments

We acknowledge funding from the European Commission of a H2020 MSCA award under proposal number 675743 (ISPIC).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2018.10.004.

References