A computational approach for phenotypic comparisons of cell populations in high-dimensional cytometry data

Ludovic Platon, David Pejoski, Guillaume Gautreau, Brice Targat, Roger Le Grand, Anne-Sophie Beignon, Nicolas Tchitchek *

CEA – Université Paris Sud 11 – INSERM U1184, Immunology of Viral Infections and Autoimmune Diseases, IDMIT Infrastructure, 92265 Fontenay-aux-Roses, France

ABSTRACT

Background: Cytometry is an experimental technique used to measure molecules expressed by cells at a single cell resolution. Recently, several technological improvements have made possible to increase greatly the number of cell markers that can be simultaneously measured. Many computational methods have been proposed to identify clusters of cells having similar phenotypes. Nevertheless, only a limited number of computational methods permits to compare the phenotypes of the cell clusters identified by different clustering approaches. These phenotypic comparisons are necessary to choose the appropriate clustering methods and settings. Because of this lack of tools, comparisons of cell cluster phenotypes are often performed manually, a highly biased and time-consuming process.

Results: We designed CytoCompare, an R package that performs comparisons between the phenotypes of cell clusters with the purpose of identifying similar and different ones, based on the distribution of marker expressions. For each phenotype comparison of two cell clusters, CytoCompare provides a distance measure as well as a p-value asserting the statistical significance of the difference. CytoCompare can import clustering results from various algorithms including SPADE, viSNE/ACCENSE, and Citrus, the most current widely used algorithms. Additionally, CytoCompare can generate parallel coordinates, parallel heatmaps, multidimensional scaling or circular graph representations to visualize easily cell cluster phenotypes and the comparison results.

Conclusions: CytoCompare is a flexible analysis pipeline for comparing the phenotypes of cell clusters identified by automatic gating algorithms in high-dimensional cytometry data. This R package is ideal for benchmarking different clustering algorithms and associated parameters. CytoCompare is freely distributed under the GPL-3 license and is available on https://github.com/tchitchek-lab/CytoCompare.

© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Flow cytometry and mass cytometry are experimental techniques measuring molecules expressed by cells, at their surface or intracellularly, at a single cell resolution [1,2]. These techniques use antibodies conjugated to fluorochromes or metals to stain cells. Cell marker expressions are then quantified using fluorescence detection systems or by time-of-flight mass spectrometry. Flow cytometry can currently measure up to 18 cell markers. Mass cytometry, which has been introduced more recently, can measure up to 40 cell markers. Both techniques offer important perspectives as they can potentially evaluate many more markers in the near future [3]. Novel bioinformatic approaches are then needed to explore and analyze complex cell interaction systems characterized by high-dimensional cytometry [4–6].

Cytometry data can be analyzed using manual or automatic gating strategies. In both strategies, the aim is to identify and quantify populations of cells, also named cell clusters, having similar phenotypes. Automatic gating methods detect cell populations using density- or expression- based approaches and produce less biased results compared to manual gating. Many automatic gating algorithms have been proposed over the last decades. Among them, SPADE [5], viSNE/ACCENSE [7,8], and Citrus [9], which are expression-based algorithms, are prevalent for high-dimensional cytometry analyses [10–15]. These algorithms use different computational strategies and are complementary to identify cell clusters with relevant biological behaviors in cytometry data.

Whereas many tools have been developed, only a limited number of computational strategies have been designed to compare the phenotypes of identified cell clusters [16–18]. However, these phe-

* Corresponding author at: CEA, 18 Route du Panorama, 92260 Fontenay-aux-Roses, France.

E-mail address: nicolas.tchitchek@cea.fr (N. Tchitchek).

notypic comparisons are crucial to explore cell clustering results obtained from different computational approaches and to recognize cell populations with common or marginal phenotypes. Thereby, comparisons are often performed manually through multiple comparisons of marker expression densities, a highly biased and time-consuming process. Comparisons of cell cluster results based on the number of shared cells among the different clusters only provide a limited set vision that cannot be used to characterize their phenotypes fully. Also, comparisons of cell cluster phenotypes based on the mean or median of marker expressions can be biased as they do not consider all the characteristics of marker expression distributions, such as the standard deviations. Proper phenotypic comparisons of cell marker expressions and cluster phenotypes must be performed based on the density distributions. Additionally, no visualization methods have been proposed to represent such phenotypic comparisons in easily interpretable ways.

To answer that need, we developed CytoCompare, an R package to compare the phenotypes of cell clusters identified by automatic gating algorithms. The aim of CytoCompare is to facilitate the easy identification of similar and different cell clusters identified by various bioinformatics approaches. For each comparison of two cell clusters, a distance measure is computed, based on the expression distribution densities of selected markers. Our approach is based on the Kolmogorov-Smirnov distance to quantify marker differences. Additionally, CytoCompare provides a p-value for determining whether two cell clusters are statistically different or not. CytoCompare proposes several visualizations methods, such as parallel coordinates, parallel heatmaps, dendrograms, multidimensional scaling (MDS) [19] or circular graphs representations [20,21], which make cell cluster phenotypes and phenotypic comparisons easily understandable and explicit. Cell clusters identified from SPADE, viSNE/ACCENSE, and Citrus algorithms can be directly imported. Importantly, we designed CytoCompare in a way that it can be easily used by non-bioinformatician experts, but can also be easily customized by users with more expertise in bioinformatics.

2. Material and methods

2.1. Availability of cytometry profiles

All the cytometry profiles used to illustrate CytoCompare are available on the FlowRepository database [22] under accession number "FR-FCM-ZZ99". This public dataset consists of 15 mass cytometry profiles of 25 cell markers obtained in the context of a macaque vaccine study, as described in [10]. Five cynomolgus macaques were vaccinated with a recombinant Modified Vaccinia virus Ankara (MVA) expressing HIV clade B antigens, which is a candidate HIV vaccine, with the aim to compare B cell responses at baseline, and 8 and 28 days after a second immunization.

2.2. Implementation in R

CytoCompare has been implemented in R. The ggplot2, ggdendro, ggrepel, grid, igraph, MASS, RJSONIO, and XML R packages are currently required for running CytoCompare. The flowCore and flowUtils Bioconductor packages are also required for running CytoCompare. The exact left-tailed binomial test, used in the statistical approach, is based on the 'binom.test()' function available in R. Comparison and intermediary results generated by CytoCompare are handled as S4 objects to allow users to access and export them easily. CytoCompare is freely distributed under the GPL-3 license and is available on https://github.com/tchitchek-lab/CytoCompare.

2.3. Graphical representations

Multidimensional scaling (MDS) and circular graph representations are generated as HTML files to allow an interactivity with users. These interactive features allow users to obtain details about the cell clusters and about the similarity associations found to be significant. The D3.js library [23] (version 3) is used by CytoCompare to create these interactive features. MDS computations are calculated using the isoMDS algorithm available in the MASS package. All other representations generated by CytoCompare are based on the ggplot library.

2.4. SPADE and viSNE/ACCENSE analyses

The SPADE analyses used to illustrate CytoCompare were performed using the publicly available SPADE R package [5]. The viSNE/ACCENSE analyses were generated using the ACCENSE software [8]. SPADE and viSNE/ACCENSE analyses were generated using different settings as detailed in the results section. Cell clusters correlating with the anti-MVA titer were identified using the SPADEVizR [4] R package, based on data published in [10].

3. Results

3.1. Definition of analysis situations requiring phenotypic comparisons of cytometry cell clusters

Cytometry experts can face at least five different situations that can necessitate comparing cell clusters based on their phenotypes.

The first situation occurs when choosing the cell clustering algorithm, which implies to compare clustering results from different algorithms. Each automatic gating algorithm is based on different computational strategies, statistical assumptions, and working hypotheses. Thus, each algorithm identifies a specific set of cell clusters with specific phenotypes. Cell clustering results obtained from different algorithms need to be compared to detect cell clusters commonly identified by various algorithms and the ones specifically identified by one algorithm. Overall, these comparisons allow to confirm clustering results with independent clustering algorithms and thus to strengthen them. These comparisons also permit to choose the automatic gating algorithm that best outlines the initial hypothesis or that provides the richest set of biological conclusions.

The second situation happens when selecting the settings for one given automatic gating algorithm, which implies to compare clustering results obtained using different parameters. For each automatic gating algorithm, different cell populations can be identified depending on the clustering settings initially defined by the user. For instance, SPADE clustering results can vary with the specified number of cell clusters to identify. Additionally, the phenotypes of the cell clusters identified by viSNE/ACCENSE can also differ depending on the perplexity and theta parameters as well as the number of iterations performed by the algorithm. Thus, the cell clustering results generated using different parameters need to be compared to unravel the effects of these parameters.

The third situation takes place when ascertaining the reproducibility of the clustering results by repeating the analysis using the same settings. Expression-based algorithms are relevant approaches to identify rare cell populations but require extensive computation capabilities. To overcome these computing difficulties, SPADE, viSNE/ACCENSE, and Citrus algorithms perform the identification of cell populations on a restricted cell subset of the input dataset. This process, named down-sampling, is usually stochastic and can lead to the identification of different cell populations through multiple runs of the algorithm. Cell clustering