Methods xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Methods for discovery and characterization of cell subsets in high dimensional mass cytometry data

Kirsten E. Diggins [a], P. Brent Ferrell Jr. [b], Jonathan M. Irish [a,c,*]

[a] Cancer Biology, Vanderbilt University School of Medicine, United States
[b] Medicine/Division of Hematology–Oncology, Vanderbilt University School of Medicine, United States
[c] Pathology, Microbiology and Immunology, Vanderbilt University School of Medicine, United States

ARTICLE INFO

Article history:
Received 16 January 2015
Received in revised form 24 April 2015
Accepted 6 May 2015
Available online xxxx

Keywords:
Mass cytometry
Flow cytometry
Single cell biology
Unsupervised analysis
Machine learning

ABSTRACT

The flood of high-dimensional data resulting from mass cytometry experiments that measure more than 40 features of individual cells has stimulated creation of new single cell computational biology tools. These tools draw on advances in the field of machine learning to capture multi-parametric relationships and reveal cells that are easily overlooked in traditional analysis. Here, we introduce a workflow for high dimensional mass cytometry data that emphasizes unsupervised approaches and visualizes data in both single cell and population level views. This workflow includes three central components that are common across mass cytometry analysis approaches: (1) distinguishing initial populations, (2) revealing cell subsets, and (3) characterizing subset features. In the implementation described here, viSNE, SPADE, and heatmaps were used sequentially to comprehensively characterize and compare healthy and malignant human tissue samples. The use of multiple methods helps provide a comprehensive view of results, and the largely unsupervised workflow facilitates automation and helps researchers avoid missing cell populations with unusual or unexpected phenotypes. Together, these methods develop a framework for future machine learning of cell identity.

1. Introduction

1.1. High dimensional single cell biology

Single cell biology is transforming our understanding of the biological mechanisms driving human diseases and healthy tissue development [1]. Mass cytometry is a recently developed technology that enables simultaneous detection of more than 40 features on individual cells [2,3]. High dimensional mass cytometry measurements are single cell, quantitative, and well-suited to unsupervised computational analysis. New analysis tools have been created to take advantage of the massive amounts of data that result from high content single cell techniques like mass cytometry. Variations of many of these tools have been developed and applied for gene expression analysis, a field facing similar problems with data dimensionality. These tools draw on advances in machine learning and statistics that are not yet widely applied in biological studies. Many of these tools are complementary and address different aspects of data analysis, and it can be challenging for biologists to know when and how to use these tools to get the most out of their data. Advances have also been made in automating and standardizing the flow cytometry data analysis workflow [4–6]. Here, we present a modular workflow focused on high dimensional single cell analysis that combines multiple tools to provide a comprehensive view of both cells and populations. Rather than making the workflow fully automated, the goal here was to combine the complementary benefits of expert analysis and machine learning. This approach maintains single cell views, provides automatic population assignment for each cell, and facilitates statistical comparison of the key cellular features that characterized each population. This semi-supervised workflow facilitates comparison of populations discovered by different computational approaches, in different clinical samples, or using different biological features (e.g. RNA expression, cell surface protein expression, and cell signaling).

An advantage of traditional analysis in flow cytometry is the reliance on identification of known, prominent populations with strong supporting biology in the literature. Given the typical panel size for fluorescent experiments, this type of supervised analysis is fast and usually adequate. Unfortunately, expert manual gating has been shown to be particularly prone to inter-operator variability [7] and a tendency to overlook cell populations [8–10]. Recent

* Corresponding author at: Vanderbilt University School of Medicine, 740B Preston Building, 2220 Pierce Avenue, Nashville, TN 37232-6840, United States.
E-mail address: jonathan.irish@vanderbilt.edu (J.M. Irish).

http://dx.doi.org/10.1016/j.ymeth.2015.05.008
1046-2023/

efforts have developed new tools for high dimensional cytometry data that bring in elements of machine learning and statistical analysis, including clustering [11–14], dimensionality reduction [8], variance maximization [15], mixture modeling [6,16–18], spectral clustering [19], neural networks [20], and density-based automated gating [21]. Here, we highlight use of these tools in a sequential single cell bioinformatics workflow (Table 1). In particular, different tools address aspects of data visualization, dimensionality reduction, population discovery, and feature comparison. It can be valuable to apply multiple tools in order to view data in different ways and fully extract biological meaning at the single cell level (Fig. 1) and the population level (Figs. 2 and 3). After identifying cell subsets with the aid of computational tools, measured features, such as protein expression in the examples here, can be compared between and within the subsets. Traditional statistics used include medians, variance, and fold changes. Other statistical methods such as histogram statistics and probability binning have also been used to compare distributions in flow cytometry data [22–24].

1.2. Overview of the analysis workflow

The workflow presented here was applied to a CyTOF dataset from the analysis of healthy human bone marrow and a diagnostic sample of blood from a patient with acute myeloid leukemia (AML). The annotated FCS files and a step-by-step guide are available online from Cytobank (www.cytobank.org/irishlab) [25] and FlowRepository (http://flowrepository.org/experiments/640) [26]. This workflow was developed for use with high-dimensional mass cytometry data. However, it can also be applied to fluorescent flow cytometry data. The main steps presented consist of event restriction, population discovery, and population characterization. Each of these aspects of data analysis can be achieved with a variety of techniques (Table 1), and some tools address multiple steps. By sequentially combining three different techniques, this workflow draws on the strengths of specific tools, keeps biologists in touch with single cell views, and enables analysis of data from different studies and single cell platforms.

In the case of the example dataset here, the overall biological goal was to identify and compare three populations of cells: leukemia cells (AML blasts) and non-malignant cells (non-blasts) in the blood of a leukemia patient, and bone marrow cells from a healthy donor. In the analysis workflow, cell events were first manually gated based on event length and DNA content to include intact, single cells (Fig. 1) [11]. Next, visualization of stochastic neighbor embedding (viSNE) was used to identify and gate major subsets (Fig. 1). Gated cells from healthy bone marrow and AML were then analyzed by spanning-tree progression analysis of density-normalized events (SPADE) to discover and compare cell subsets (Fig. 2). Finally, the cell subsets identified by SPADE were further characterized using complete linkage hierarchical clustering and a heatmap in R (Fig. 3). The details of mass cytometry data collection and processing prior to initial cell selection (gating) are not covered in detail here. These early steps include experiment design, collection of data at the instrument (and instrument setup), any normalization, and transformation of the data to an appropriate scale (Table 1).

The initial event restriction step that begins the workflow focuses the analysis on populations of cells. The goal at this step is to remove events that do not contribute useful information while making minimal changes to the data and not over-focusing. Event restriction is traditionally performed using biaxial gating (Table 1), but given the high dimensionality of mass cytometry data, use of viSNE (Fig. 1) can simplify the process of distinguishing initial populations and avoid overlooking cells with unusual or unanticipated phenotypes. The second step, cell subset identification, is also traditionally performed by expert gating (Table 1). However, clustering tools such as SPADE [12] (Fig. 2), Misty Mountain [13], and Citrus [14], among others, can be used to automatically assign cells to groups or clusters in high dimensional data. In the workflow here, the goal is to find all the phenotypic clusters of cells in healthy bone marrow, AML blasts, and non-blast cells from AML blood (Fig. 2). As the final step, characterization of discovered cell

Table 1
A modular machine learning workflow for semi-supervised high-dimensional single cell data analysis.

Analysis step		Traditional	Additional methods[§]	Method here
Data collection	(1) Panel design	Human expert	–	–
	(2) Data collection	Human expert	–	–
Data processing	(3) Cell event parsing	Instrument software	Bead normalization and event parsing [39]	–
	(4) Scale transformation	Human expert	Logicle [47]	–
Distinguishing initial populations	(5) Live single cell gating (6) Focal population gating	Biaxial gating + human expert	No event restriction, AutoGate [61]	viSNE + human expert (Fig. 1)[†]
Revealing cell subsets	(7) Select features	Human expert	Statistical threshold [53]	Human expert[†]
	(8) Reduce dimensions or transform data	N/A	Heat plots [62], SPADE [12], *t*-SNE [63], viSNE [8], ISOMAP [27], LLE [29], PCA in R/flowCore [64]	SPADE[†], viSNE
	(9) Identify clusters of cells	Human expert	SPADE, *k*-medians, R/flowCore, flowSOM [65], Misty Mountain [13], JCM [30], ACCSENSE [66], DensVM [28], AutoGate, *Citrus* [14]	SPADE (Fig. 2)[†], viSNE + human expert (Fig. 1)
	(10) Cluster refinement	Human expert	Citrus, DensVM, R/flowCore	–
Characterizing cell subsets	(11) Feature comparison	Select biaxial single cell views	viSNE, SPADE, heatmaps [25,53], histogram overlays [25,53], violin or box and whiskers plots [64], wanderlust [31], gemstone	Heatmaps (Fig. 3A)[†], viSNE (Fig. 1C), SPADE (Fig. 2C)
	(12) Model populations	N/A	Median [53], JCM, PCA	–
	(13) Learn cell identity	Human expert	–	Human expert[†] (Figs. 1B, 2B, and 3B)
	(14) Statistical testing	Prism, excel	R/flowCore	–

[§] Methods with broad application (e.g. R/flowCore) are listed minimally at select steps based on particular strengths or published applications.
[†] Denotes the primary approach used at each step in the sequential analysis workflow shown here.