TEAM: A Multiple Testing Algorithm on the Aggregation Tree for Flow Cytometry Analysis

John Pura\textsuperscript{a}, Xuechan Li\textsuperscript{b}, Cliburn Chan\textsuperscript{b}, Jichun Xie\textsuperscript{b,\ast}

\textsuperscript{a} Center of Innovation to Accelerate Discovery and Practice Transformation, Durham Veterans Affairs Medical Center, Durham, NC 27701
\textsuperscript{b} Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27705

Abstract

In immunology studies, flow cytometry is a commonly used multivariate single-cell assay. One key goal in flow cytometry analysis is to detect the immune cells responsive to certain stimuli. Statistically, this problem can be translated into comparing two protein expression probability density functions (PDFs) before and after the stimulus; the goal is to pinpoint the regions where these two PDFs differ. Further screening of these differential regions can be performed to identify enriched sets of responsive cells. In this paper, we model identifying differential density regions as a multiple testing problem. First, we partition the sample space into small bins. In each bin, we form a hypothesis to test the existence of differential PDFs. Second, we develop a novel multiple testing method, called TEAM (Testing on the Aggregation tree Method), to identify those bins that harbor differential PDFs while controlling the false discovery rate (FDR) under the desired level. TEAM embeds the testing procedure into an aggregation tree to test from fine- to coarse-resolution. The procedure achieves the statistical goal of pinpointing density differences to the smallest possible regions. TEAM is computationally efficient, capable of analyzing large flow cytometry data sets in much shorter time compared with competing methods. We applied TEAM and competing methods on a flow cytometry data set to identify T cells responsive to the cytomegalovirus (CMV)-pp65 antigen stimulation. With additional downstream screening, TEAM successfully identified enriched sets containing monofunctional, bifunctional, and polyfunctional T cells. Competing methods either did not finish in a reasonable time frame or provided less interpretable results. Numerical simulations and theoretical justifications demonstrate that TEAM has asymptotically valid, powerful, and robust performance. Overall, TEAM is a computationally efficient and statistically powerful algorithm that can yield meaningful biological insights in flow cytometry studies.

Keywords: Flow cytometry, Multiple testing, aggregation tree, distribution difference, false discovery proportion (FDP)

\*Principal corresponding author

Email addresses: john.pura@duke.edu (John Pura), xuechan.li@duke.edu (Xuechan Li), cliburn.chan@duke.edu (Cliburn Chan), jichun.xie@duke.edu (Jichun Xie)
1. Introduction

Flow cytometry is a multivariate single-cell assay commonly used to characterize the immune system. A key challenge in flow cytometry is the identification of T cells that are activated by specific antigens such as a particular tumor, bacterial or viral protein. Intracellular cytokine staining (ICS) is often combined with flow cytometry to analyze the antigen-specific T cell immune response. In ICS, cells are first activated with antigen, followed by staining for cell surface molecules that define the cell phenotype, such as CD4 and/or CD8. Cells can be further stained after membrane permeabilization with fluorochrome-labeled monoclonal antibodies specific to protein markers within the cell, such as the IFN-γ, IL-2, and TNF-α cytokines that are only expressed after T cell activation. Hence T cells with high levels of cytokine expressions are likely to be antigen-specific. The flow cytometer quantifies the amount of antibodies bound to the cell and hence the concentration of the protein marker that the antibody is targeting. The expression repertoire of protein markers is often used to characterize cell types at two levels – the cell surface markers, which define the basic cell type (e.g., CD4+ T cell) and maturational stage (e.g., memory T cells), and the intracellular cytokines, whose expressions define a cell’s activation class. Quantification of different activation classes of antigen-specific T cells provides useful biological information such as the likely efficacy of a vaccine (Seder et al., 2008). However, as the relative frequency of antigen-specific cells for any particular antigen is often very low (rarely above 1% and often much lower), several hundred thousand cells per sample are typically evaluated to quantify different subpopulations of antigen-specific T cells. This gives rise to the need for statistical methods that can detect antigen-specific cells but guard against false positives.

As an illustrative example, we present a dataset from the External Quality Assurance Program Oversight Laboratory proficiency program by the Duke Immune Profiling Core (Staats et al., 2014). Blood samples from 11 healthy individuals were collected, and each sample was split into two parts. One was used as a negative control (cohort 1); the other was stimulated with a peptide mixture from the immunodominant cytomegalovirus pp65 protein (cohort 2). Cohort 1 contains nearly 2.4 million cells and cohort 2 has nearly 2.2 million cells. Each cell has 11 protein markers, collected for discriminating among T cell basic, maturational and functional subsets. We expect that specific T cells targeting the cytomegalovirus would be activated and show elevated expression levels in the four functional protein markers: TNF-α, IL-2, IFN-γ, and CD107 (a protein associated with cytotoxic activity). The one- and two-dimensional marker density contour plots are presented in Figure 1.

Generally speaking, in a typical flow cytometry study, \( N_1 \) cells under condition 1 (cohort 1) and \( N_2 \) cells under condition 2 (cohort 2) are evaluated; for each cell, \( p \) protein marker expressions are collected. Here \( N_1 \) and \( N_2 \) could range from \( 10^5 \) to \( 10^7 \), and \( p \) could range from fewer than 10 up to 50. The goal is to pinpoint the activated cells. The characterization of cells activated under condition 2 has many applications - for example, in evaluating the immunogenicity of a vaccine. As mentioned above, the activated cells are typically present in very low relative frequencies. Identifying those cells directly using an analytic method is very challenging. Thus, we first identify the sample space regions that are highly enriched
Figure 1: Univariate and bivariate distribution of activation markers on cells evaluated by flow cytometry. Potential biological differences between cohort 1 (control; solid line) and cohort 2 (stimulated; dashed line) samples can be identified through examining bivariate densities of the four functional markers: TNF-α, IL-2, IFN-γ, and CD107. Univariate densities are plotted along the diagonals, while bivariate densities are plotted in the lower triangular region. Cohort 1 functional marker densities are displayed as greyscale contour lines, while cohort 2 marker densities are displayed as filled, greyscale contours.

for such cells. These regions could still contain some non-activated cells, so further filtering is needed to pinpoint the activated cells. The filtering could be based on other cell protein markers from the domain knowledge or based on sub-analyses of multiple functional markers. See details in Section 4.

Many existing methods used this strategy to first identify differential density regions. Roederer and Hardy (2001) proposes the frequency gating method (Roederer et al., 2001) to compare the frequencies between stimulated and reference cells and identify regions with significant differential frequencies. The reference sample space is divided into bins containing roughly equal numbers of cells. The resulting partition is then applied to the stimulated sample, and the ratio of the stimulated and reference cells in each bin is calculated to derive a normalized chi-square statistic for each bin. Then they use a user-defined threshold to reject those bins with large chi-square statistics. Duong (2013) uses kernel density estimation and chi-square test statistics to identify local distributional differences at any given location in multi-dimensional space. Antoniadis et al. (2015) focuses on identifying one-dimensional differential density regions. They divide the samples into equal length bins and then use the Poisson distribution to model the number of stimulated and reference cells in each bin.
Then they used a variance stabilization transformation to normalize the counts and applied existing multiple-testing procedures to identify differential density regions. More recently, Soriano and Ma (2017) proposed a multi-resolution scanning (MRS) method for identifying differential density regions. The method can be viewed as a testing method embedded in the partition tree. MRS sequentially partitions the common support of two distributions into finer and finer bins. On every resolution level, each bin is coupled with a null hypothesis. MRS forms a hypothesis for each large or small bin and asymptotically controls the false discovery rate (FDR) among all these hypotheses.

In this paper, we model the challenge of pinpointing differential regions as a multiple-testing problem, and propose a new FDR controlling procedure, called TEAM. First, TEAM partitions the multivariate sample space into bins with the finest resolution. It can accommodate different partitioning schemes. Second, TEAM embeds testing on an aggregation tree. On layer 1, within each bin, TEAM tests if the pdf of cohort 2 is higher than that of cohort 1. On higher layers, TEAM will gradually aggregate the accepted bins and test if the aggregated bins harbor differential pdfs. This fine-resolution to coarse-resolution testing structure not only boosts the testing power but also pinpoints the regions with differential pdfs at the finest possible resolution.

Although both TEAM and MRS are both testing methods embedded in a generated tree, there are several fundamental differences between these two methods. First, TEAM is embedded in an aggregation tree. It starts bottom-up from the finest bins for testing and gradually aggregates these bins to borrow information across local regions. In contrast, MRS is embedded in the partition tree. It starts top-down from the coarsest bins for testing and gradually partitions the bins to identify local differences. Second, TEAM prioritizes the identification of small bins and pinpoints the regions harboring differential pdfs at the finest possible resolutions. In contrast, MRS tends to prioritize global differences. For example, consider the situation where cohort 2 density is higher than cohort 1 in a very small region. Because MRS starts from large bins if the large bin contains some small differential density region, the signal in the small region could be masked by the noise in the large bin. As a result, MRS will accept the entire region corresponding to the large bin in the early stages. Furthermore, MRS will prune some accepted regions in the early stages to improve computation efficiency. If the accepted large bin is pruned, the small differential region will never be identified in later stages. Third, for FDR control, TEAM only considers the null hypotheses coupled with the finest-resolution bins, which by design are non-overlapping. These fine-resolution bins stand alone and are free to be null or alternative by themselves. In contrast, MRS defines FDR based on all hypotheses coupled with bins at any resolution level so that higher-resolution bins are nested in the lower-resolution bins. In MRS, it is possible to reject a parent bin but accept all its child bins; it is also possible to accept a parent bin but reject all its child bins. Because all parents and child hypotheses are counted as candidate hypotheses, interpreting the conflicting results is challenging. In TEAM, accepting child nodes while rejecting the parent nodes could also happen because the child nodes’ signal-to-noise ratio could be too small to claim rejection. However, after aggregation, the parent node’s signal-to-noise ratio is large enough to claim rejection. However, this will not cause any ambiguity in interpretation because all candidate hypotheses are defined only based on
the finest-resolution nodes (leaves): a leaf could be either rejected or accepted.

Besides the MRS method, a number of hierarchical testing methods have been developed in recent years. Dmitrienko and Tamhane (2011) and Dmitrienko and Tamhane (2013) developed multiple testing procedures to test the hypotheses, which are grouped into hierarchically ordered families and control the family-wise error rate control (FWER). The methods have applications in clinical trials. Goeman and Finos (2012) and Meijer and Goeman (2015) developed the FWER controlling procedures on the trees and directed acyclic graphs (DAG). Some other methods (Yekutieli, 2008; Guo et al., 2018; Ramdas et al., 2019; Bogomolov et al., 2020; Li et al., 2020) are developed to control node-level FDR, and these methods usually assume certain graphical structures exist among these node hypotheses. The settings are very different from our paper aiming to find differential density regions; thus the existing methods cannot be applied to our problem.

The rest of the paper is organized as follows. Section 2 introduces how to model the flow cytometry data. Section 3 describes the TEAM algorithm. Section 4 focuses on analyzing the flow cytometry data to identify cells responsive to the cytomegalovirus antigen. Section 5 discusses several numerical settings and compares the performance of TEAM and MRS. Section 6 provides a summary of this paper and a general discussion on the application of TEAM to other related single-cell assays.

2. Model

2.1. Sample space partitioning

Let $\Omega$ be the multivariate protein marker sample space of the pooled cells (cohort 1 and cohort 2). Consider a partition on $\Omega$, i.e., $\bigcup_{i=1}^{m} \Omega_i = \Omega$, $\mu(\Omega_i \cap \Omega_j) = 0$ if $i \neq j$. We call $\Omega_1, \ldots, \Omega_m$ the leaf bins (of leaf 1, $\ldots$, m). Suppose the leaf bin $\Omega_i$ consists of $n_i$ cells. The partition can be constructed in several ways. Two examples include:

- **Adaptive partition.** We order the protein markers from the largest expression variance to the smallest. We choose the marker with the largest sample variance and do a median split of the sample space along this dimension. Within each subsample, we repeat these median splits of the sample space along the dimension with the largest variance. After $\tilde{m}$ times, we will have $m = 2^\tilde{m}$ bins. See ? and Roederer et al. (2001) for details.

- **Sequential partition.** We order the protein markers from the largest expression variance to the smallest. We first partition the first marker dimension into $\tilde{m}$ bins by its sample quantiles at level $\{1/\tilde{m}\}, \ldots, \{(\tilde{m}-1)/\tilde{m}\}$. Within each bin on the first dimension, we partition the second dimension by its sample quantiles at level $\{1/\tilde{m}\}, \ldots, \{(\tilde{m}-1)/\tilde{m}\}$. We then sequentially partition all other dimensions until all dimensions are partitioned. Please see Figure 3 for an illustration.

2.2. Hypotheses

Let $\hat{X}_i$ and $X_i$ be the number of cells from cohort 1 and cohort 2 that falls into leaf bin $i$, respectively. Then $n_i = X_i + \hat{X}_i$. We know that $\sum_{i=1}^{m} \hat{X}_i = N_1$ and $\sum_{i=1}^{m} X_i = N_2$. Consider
theoretical choice of partitioned bins, the goal is to identify like to identify where can first run the one-sided test, and flip the labels of cohort 1 and cohort 2, and run the

differential densities in either direction; then we need to find regions where the cohort 2 density is higher. Under some rare cases, researchers

continuity of $H$. If $H = \text{null}$ we call leaf $i$ alternative and null otherwise.

\begin{equation}
\theta_i = \frac{N_2 \int_{\Omega_1} f_2(y)dy}{N_2 \int_{\Omega_2} f_2(y)dy + N_1 \int_{\Omega_1} f_1(y)dy}
\tag{1}
\end{equation}

Here, $f_1$ and $f_2$ are the PDFs of cohort 1 and cohort 2, and $N_1$ and $N_2$ are the cell numbers in cohort 1 and cohort 2.

From here, we consider this problem in the probability space after partition. Noteworthy, the new sample space after partition $\tilde{\Omega}$ contains all possible realizations of $X_1, \ldots, X_m$

conditioning on $N_1, N_2$, and $n_1, \ldots, n_m$. It is different from the original sample space $\Omega$. We also need to define the probability measure conditioning on the partition. Lemma 1 characterizes this measure.

**Lemma 1.** Let $Z_1, \ldots, Z_m$ be a sequence of independent random variables, where $Z_i$ follows Binom($n_i, \theta_i$). Suppose $n^{(1)} = \sup_{i=1}^m n_i = n_i + \delta_{n,i}$ with $\sup_{i=1}^m |\delta_{n,i}| = o(n^{(1)})$ and $n^{(1)} = o(N^{1/2})$. For any constant $K$ and vector $(i_1, \ldots, i_K)$, with each element taken without replacement from $\{1, \ldots, m\}$,

\[
\left| \frac{P(X_{i_1} = x_{i_1}, \ldots, X_{i_K} = x_{i_K} \mid N_1, N_2, n_{i_1}, \ldots, n_{i_K})}{\prod_{j=1}^K P(Z_{i_j} = x_{i_j} \mid n_{i_j})} - 1 \right| \leq CK^2(n^{(1)})^2/N,
\]

for some constant $C$ not depending on $K$, $n^{(1)}$, $m$, $N_1$, $N_2$, or $(x_{i_1}, \ldots, x_{i_K})$.

When we justify the asymptotic properties of TEAM, only joint distributions of finite dimensional $(X_{i_1}, \ldots, X_{i_K})'$ are involved. Therefore, Lemma 1 is sufficient.

Let $\theta_0 = N_2/N$. For leaf $i$, we set up the hypothesis:

\begin{equation}
H_{\text{null},i} : \theta_i \leq \theta_0 \quad \text{versus} \quad H_{\text{alt},i} : \theta_i > \theta_0.
\tag{2}
\end{equation}

If $H_{\text{alt},i}$ is true we call leaf $i$ alternative and null otherwise.

We consider one-sided tests here because they correspond to our analytical goal of locating the activated cells in cohort 2. Clearly, $\int f_s(y)dy = 1$ for both $s \in \{1, 2\}$. By the

continuity of $f_1$ and $f_2$, if there exists a region where the cohort 2 density is higher, there must exist a region where the cohort 1 density is higher. For our analytical goal, we only need to find regions where the cohort 2 density is higher. Under some rare cases, researchers are interested in identifying regions with differential densities in either direction; then we can first run the one-sided test, and flip the labels of cohort 1 and cohort 2, and run the one-sided test again.

Let $\mathcal{H} = \{1, \ldots, m\}$, $\mathcal{H}_{\text{null}} = \{i : H_{\text{null},i} \text{ is true}\}$, and $\mathcal{H}_{\text{alt}} = \mathcal{H} \setminus \mathcal{H}_{\text{null}}$. Ideally, we would like to identify where $f_2 > f_1$, i.e., the region $\Omega^+ = \{y \in \Omega : f_2(y) > f_1(y)\}$. With the partitioned bins, the goal is to identify $\hat{\Omega}^+ = \bigcup_{i \in \mathcal{H}_{\text{alt}}} \Omega_i$. To make sure that $\hat{\Omega}^+$ approximates $\Omega^+$ well, the partition should be fine enough. However, if the partition is too fine and each leaf bin only contains very few cells, the testing power will be low. Discusssions on the proper theoretical choice of $m$ and $n_i$ are provided in the supplementary file.
3. Algorithm Description

3.1. Step 1: Testing on layer 1

The false discovery proportion (FDP) and false discovery rate (FDR) is defined as

\[ \text{FDP} = \frac{\sum_{i \in H_{\text{nul}}} I(H_{\text{nul},i} \text{ is rejected})}{\sum_{i \in H} I(H_{\text{nul},i} \text{ is rejected})} \land 1, \quad \text{FDR} = \mathbb{E}(\text{FDP}). \] (3)

To control FDR, we propose the following test procedure on the bottom layer. Let \( G^{(1)}_{0,i} \) be the complementary cumulative density function (ccdf) of Binom\((n_i, \theta_0)\). Let \( P^{(1)}_{0,i} = G^{(1)}_{0,i}(X_i) \), a random variable close to the P-value. Define the threshold \( \hat{c}^{(1)} \) as

\[ \hat{c}^{(1)} = \sup \left\{ a^{(1)}_N \leq c \leq \alpha : c = \frac{\max \left\{ \sum_{i \in H} I(P^{(1)}_{0,i} < c) \right\}}{m} \cdot \alpha \right\}. \] (4)

Here \( a^{(1)}_N = (m \log m)^{-1} \). If such \( \hat{c}^{(1)} \) does not exist, set \( \hat{c}^{(1)} = a^{(1)}_N \). We reject \( H_{\text{nul},i} \) if \( P_{0,i} \leq \hat{c}^{(1)} \). This procedure is very similar to the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995).

3.2. Step 2: Aggregation and Testing on higher layers.

Unlike other multiple testing methods, TEAM will continue testing after layer 1. It will aggregate the neighboring accepted leaves to parent nodes and test their parent hypotheses. The underlying assumption of TEAM is that the neighboring leaves of an alternative leaf are also likely to be alternative. This assumption is reasonable under many circumstances, especially when the PDFs \( f_1 \) and \( f_2 \) are smooth (See Proposition 1). Thus, TEAM hierarchically aggregates the neighboring leaves so that their signals can be aggregated and amplified. It is possible that an alternative region spanning multiple leaf bins is missed on low layers but will be identified on higher layers.

Figure 2 provides a toy example to illustrate how TEAM works.

- Leaf 1 and leaf 2 are accepted on layer 1, so they are aggregated as a parent node \( S^{(2)}_1 = \{1, 2\} \). The coupled node hypothesis is

\[ H^{(2)}_{\text{nul},1} : \forall j \in S^{(2)}_1, \theta_j \leq \theta_0 \quad \text{versus} \quad H^{(2)}_{\text{alt},1} : \exists j \in S^{(2)}_1, \theta_j > \theta_0. \]

- Leaf hypothesis \( H^{(1)}_{\text{nul},8} \) is rejected on the bottom layer. As a result, leaves 7 and 9 are aggregated on layer 2, so that \( S^{(2)}_4 = \{7, 9\} \). This aggregation design allows the differential peak to be captured at a lower layer and the differential shoulder to be captured at a higher layer. See the illustrated distributions in Figure 2. Aggregating the shoulder areas will likely increase power.

- Leaf 12 is left alone on the bottom layer because no more leaves are left to be aggregated. On any layer, at most 1 node will be left out.
TEAM with three layers

Final rejection set: \( R = \{1, 2, 3, 4, 7, 8, 9\} \)

\[
R^{(3)} = \{ j : j \in S_1^{(3)}, p_{0,j}^{(3)} < \hat{c}^{(3)} \} = \{1, 2, 3\}
\]

\[
R^{(2)} = \{ j : j \in S_1^{(2)}, p_{0,j}^{(2)} < \hat{c}^{(2)} \} = \{7, 9\}
\]

\[
R^{(1)} = \{ i : p_i^{(1)} < \hat{c}^{(1)} \} = \{8\}
\]

Figure 2: An illustrating example of TEAM with three layers. The non-rejected bins are aggregated at the beginning of layer 2 and layer 3, and each parent bin is coupled with a parent hypothesis. If a parent hypothesis is rejected, the rejection is mapped back to the bottom layer. For example, at the beginning of layer 2, the non-rejected leaf bin set is \( \tilde{H}^{(2)} = \{1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12\} \), and the parent bin set is \( A^{(2)} = \{\{1, 2\}, \{3, 4\}, \{5, 6\}, \{7, 9\}, \{10, 11\}\} \). On layer 2, the null hypothesis coupled with the parent bin \{5, 6\} is rejected. The rejection is mapped to the bottom layer so that \( H_{null,5} \) and \( H_{null,6} \) are rejected.

More generally, higher layers of TEAM employ two steps: aggregation and testing.

**Aggregation.** On layer \( \ell \), we aggregate the neighboring accepted child nodes on layer \( \ell - 1 \) into the parent nodes. If the leaf bins are ordinal, the aggregation can be easily performed according to the ordinal rankings. Figure 2 illustrates a one-dimensional example, and Figure 3 illustrates a two-dimensional example. The two-dimensional example can be easily extended to multiple dimensions.

**Node hypothesis.** After aggregation, new parent nodes \( S_{i}^{(\ell)} \) are formulated. We set up the coupled node hypothesis:

\[
H_{null,i}^{(\ell)} : \forall j \in S_{i}^{(\ell)}, \theta_j \leq \theta_0 \quad \text{versus} \quad H_{alt,i}^{(\ell)} : \exists j \in S_{i}^{(\ell)}, \theta_j > \theta_0.
\]  

(5)

We test these hypotheses on layer \( \ell \) and map the rejections back to the bottom layer with the finest resolution.

**Testing.** On layer \( \ell \), suppose there are \( m^{(\ell)} \) aggregated nodes on layer \( \ell \). For \( \ell \geq 2 \), each node \( S_{i}^{(\ell)} \) is the union of two child nodes on \( \ell - 1 \), denoted by \( S_{i1}^{(\ell-1)} \) and \( S_{i2}^{(\ell-1)} \). Obviously \( \text{Card}(S_{i}^{(\ell)}) = 2^{\ell-1} \). The node bin contains \( n_{i}^{(\ell)} = \sum_{j \in S_{i}^{(\ell)}} n_{j}^{(1)} \) samples, out of which \( X_{i}^{(\ell)} = \sum_{j \in S_{i}^{(\ell)}} X_{j}^{(1)} \) are from cohort 2. It is easy to see that \( X_{i}^{(\ell)} = X_{i1}^{(\ell-1)} + X_{i2}^{(\ell-1)} \), where \( X_{i1}^{(\ell-1)} \) and \( X_{i2}^{(\ell-1)} \) are the number of cohort 2 samples in the bins of its child node.
Similar to layer 1, we first derive a P-value-like statistic \( P_{0,i}^{(\ell)} = G_{0,i}^{(\ell)}(X_i^{(\ell)}; \hat{c}^{(\ell-1)}) \). Here \( G_{0,i}^{(\ell)}(c; \hat{c}^{(\ell-1)}) \) is defined recursively. On layer 1, \( G_{0,i}(c; \hat{c}(0)) \) is the CDF of Binom\((n_i^{(1)}, \theta_0)\). On layer \( \ell \geq 2 \), define

\[
G_{0,i}^{(\ell)}(c; \hat{c}^{(\ell-1)}) = P_{\theta_0}(Z_1 > c \mid G_{0,i_1}^{(\ell-1)}(Z_1; \hat{c}^{(\ell-2)}) > \hat{c}^{(\ell-1)}, G_{0,i_2}^{(\ell-1)}(Z_2; \hat{c}^{(\ell-2)}) > \hat{c}^{(\ell-1)}),
\]

where \( Z_1 \) and \( Z_2 \) independently follow Binom\((n_i^{(\ell-1)}, \theta_0)\) and Binom\((n_i^{(\ell-1)}, \theta_0)\) respectively, and \( Z = Z_1 + Z_2 \).

Technically speaking, \( P_{0,i}^{(\ell)} \) is not the P-value of \( X_i^{(\ell)} \) because the null distribution of \( X_i^{(\ell)} \) depends on the entire testing path. However, we can show that \( |P_{0,i}^{(\ell)} - P_{0,i}|/\hat{P}_{0,i} \) converges to zero in probability, where \( \hat{P}_{0,i} \) is the P-value of \( X_i^{(\ell)} \) (Lemma 8). Thus \( P_{0,i}^{(\ell)} \) asymptotically follows Unif\((0, 1)\) under \( H_{null}^{(\ell)} \).

We will reject the node hypothesis \( H_{null,i}^{(\ell)} \) if \( P_{0,i}^{(\ell)} \leq \hat{c}^{(\ell)}(\alpha) \), where

\[
\hat{c}^{(\ell)}(\alpha) = \sup \left\{ c : \alpha - \sum_{1 \leq i \leq m(i)} P_{0,i}^{(\ell)} \leq \max \left[ \frac{\sum_{1 \leq i \leq m(i)} I\{P_{0,i}^{(\ell)} \leq c\}}{m(i)} \right], \alpha \right\}
\]

with \( a_N^{(\ell)} = (m(i) \log m(i))^{-1} \). If such \( \hat{c}^{(\ell)} \) does not exist, set \( \hat{c}^{(\ell)} = a_N^{(\ell)} \).

Here, \( a_N^{(\ell)} \) is set at \( (m(i) \log m(i))^{-1} \) because later we will prove

\[
P\left( \sup_{i \in B_{null}^{(\ell)}} P_{0,i}^{(\ell)} < a_N^{(\ell)} \right) \leq \sum_{i \in B_{null}^{(\ell)}} P(P_{0,i}^{(\ell)} < a_N^{(\ell)}) = O(1/\log m) = o(1),
\]

where \( B_{null}^{(\ell)} \) is the null node set on layer \( \ell \). In other words, if the threshold is set up at \( \hat{c}^{(\ell)} \), the probability of making any false rejections on layer \( \ell \) is negligible. Thus, there is no need to consider smaller cutoffs. Similar bounds are also used in Liu (2013), Liu et al. (2014), Xie and Li (2018).

To map the node-level rejections to the leaves (on layer 1), we adopt an aggressive approach:

If \( H_{null,i}^{(\ell)} \) is rejected, then \( \forall j \in S_i^{(\ell)}, H_{null,j} \) are rejected.

For example, in Figure 2, we reject \( H_{null,1}^{(3)} \). Its corresponding node is \( S_i^{(3)} = \{1, 2, 3, 4\} \). Then we reject \( H_{null,1}, H_{null,2}, H_{null,3}, \) and \( H_{null,4} \). This aggressive approach is based on the underlying assumption that the alternative leaves are likely to cluster. See Section S1 in the supplementary file for justification.

The hierarchical structure of TEAM is designed to boost the testing power. On the bottom layer, we test one leaf at a time. When a leaf harbors a strong signal \( \theta_i \), it is likely to be rejected on layer 1. On higher layers, leaves are aggregated into larger nodes, and these nodes will be tested collectively. Because the neighboring leaves all have similar signal levels \( \theta_i \) (based on Condition C4 in the supplementary materials), some weak-signal leaves
have the chance to be aggregated. Therefore, the collective signal of the node will be much stronger, and the node will have a higher chance to be rejected. The rejections will be mapped back to the leaf layer. Thus, compared to just testing the individual leaves on the bottom layer, TEAM will be more powerful.

3.3. Stopping rule
To control when TEAM stops, we set up a flag: flag = 0 for proceeding and flag = 1 for stopping. At the beginning of TEAM, flag = 0 and switches to 1 when the stopping rule is satisfied. Here are some examples of stopping rules.

- We set up a predetermined number $L$ as the maximum layer. Then TEAM will stop after $L$ layers.
- After the testing procedure on layer $\ell$ for any $\ell \geq 2$, we calculate the number of rejections on layer $\ell$. If the number is less than a prespecified level, TEAM will stop.
- After the testing procedure on layer $\ell$ for any $\ell \geq 2$, we calculate the ratio between the rejection number on layer $\ell$ and on layer $\ell - 1$. If the ratio is below a prespecified level, TEAM will stop.

3.4. Pseudocode for TEAM
1. Set flag = 0 and $\ell = 1$.
2. On layer 1, define the leaf $S^{(1)}_i = \{i\}$ for $i = 1, \ldots, m^{(1)}$ with $m^{(1)} = m$. Let the rejection set be
   \[ R^{(1)} = \{i : P^{(1)}_{0,i} \leq \hat{c}^{(1)}\} \]
   with $\hat{c}^{(1)}$ defined in (4).
3. Check the stopping rule. If it is satisfied, set flag = 1 and go to Step 4; otherwise, increase $\ell$ by 1 and perform the following sub-steps on layer $\ell$ ($\ell \geq 2$).
   (a) Let $\tilde{\mathcal{H}}^{(\ell)} = \mathcal{H}^{(\ell-1)} \setminus R^{(\ell-1)}$.
   (b) Based on the predefined aggregation rule, let the node $S^{(\ell)}_i = S^{(\ell-1)}_{i_1} \cup S^{(\ell-1)}_{i_2}$ for $i = 1, \ldots, m^{(\ell)}$. Here $\text{Card}(S^{(\ell)}_i) = 2^\ell - 1$ and $m^{(\ell)} = \lceil \text{Card}(\tilde{\mathcal{H}}^{(\ell)})/2^{\ell-1} \rceil$.
   (c) Set $\mathcal{H}^{(\ell)} = \bigcup_{i=1}^{m^{(\ell)}} S^{(\ell)}_i$, $\mathcal{H}^{(\ell)}_{\text{mul}} = \mathcal{H}^{(\ell)} \cap \mathcal{H}^{(\ell)}_{\text{mul}}$, and $\mathcal{H}^{(\ell)}_{\text{alt}} = \mathcal{H}^{(\ell)} \cap \mathcal{H}^{(\ell)}_{\text{alt}}$.
   (d) Obtain the rejection set
   \[ R^{(\ell)} = \{j : j \in S^{(\ell)}_i, P^{(\ell)}_{0,i} \leq \hat{c}^{(\ell)}\} \]
   with $\hat{c}^{(\ell)}$ defined in (7).
4. Let the overall rejection set be $\mathcal{R} = \bigcup_{\ell=1}^{L} R^{(\ell)}$. We reject $\mathcal{H}^{(\ell)}_{\text{mul},i}$ for all $i \in \mathcal{R}$. 

10
Figure 3: An example of sequential partitioning and aggregation in a two-dimensional sample space. We partition the sample space into 64 bins to facilitate visualization. (a) Partition the two-dimensional sample space using the sequential partition strategy. Then all the bins are sequentially ordered. (b) On layer 1, the rejection set is $R^{(1)} = \{2, 10, 53, 55\}$. (c) On layer 2, the remaining bins are aggregated with their neighbors along the order defined by sequential partition. The indices of each bin in layer 2 are relabeled and comprise the bin set $S^{(2)}$. After the testing procedure, null hypotheses coupled with bin set 20 and 28 are rejected. After mapping the rejections to layer 1, the rejection set is $R^{(2)} = \{41, 42, 59, 60\}$. (d) On layer 3, we further aggregate the neighboring bin sets. After the testing procedure on layer 3, we reject the null hypotheses coupled with bin set 13. After mapping the rejections to layer 1, the rejection set is $R^{(3)} = \{54, 56, 57, 58\}$. After three layers, the overall rejection set is $R = \{2, 10, 53, 54, 55, 56, 57, 58, 59, 60\}$.
4. Application to Antigen-Specific, T cell Activation in EQAPOL Data

4.1. Preprocessing

The dataset description is provided in the Introduction (Section 1). The dataset was preprocessed using manual gating in FlowJo software (v9.9.6) to remove debris, doublet, and aggregate cells. To address the between-sample variation across the 11 individuals, we applied the quantile normalization on a per-channel basis (Hahne et al. (2010b)) to each negative control (cohort 1) and the corresponding stimulated sample (cohort 2) before pooling.

4.2. Monofunctional, bifunctional, and polyfunctional cell-enriched sets

TEAM aims to identify the differential density regions where the cohort 2 cells are more abundant than the cohort 1 cells. It cannot directly pinpoint the activated cells. Therefore, we need a careful design to transform the differential density regions into the sets that are enriched for the activated cells. The main idea is to utilize the subspaces spanned by markers.

In the EQAPOL data, we expect the activated cells to differentially express in four functional markers: TNF-α, IL2, IFN-γ, and CD107. We consider six bivariate subspaces spanned by these markers, denoted by \((\text{TNF-}\alpha, \text{IL2})\), \((\text{TNF-}\alpha, \text{IFN-}\gamma)\), \((\text{TNF-}\alpha, \text{CD107})\), \((\text{IL2}, \text{IFN-}\gamma)\), \((\text{IL-2}, \text{CD107})\), and \((\text{IFN-}\gamma, \text{CD107})\).

Consider the subspace with spanned by marker \(i\) and marker \(j\), \(i < j\), and denote the true differentially density region in this subspace by \(\Omega_{ij}^+\). Let \(Y_{ci}\) be the expression of cell \(c\) and marker \(i\). Define the rejection cell sets

\[ C_{ij} = \{c : (Y_{ci}, Y_{cj}) \in \Omega_{ij}^+\}, \]

with

\[(i, j) \in \mathcal{S} = \{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)\}.\]

Although \(C_{ij}\) could contain both activated and non-activated cells, by taking intersections of these sets, we can further narrow down those cells to mostly activated cells. These activated cells can be further divided into three types: monofunctional, bifunctional, and polyfunctional. For any cell, depending on its cell type (non-functiona, monofunctional, bifunctional, and polyfunctional), we expect it will falls in different sets represented by \(C_{ij}\)s.

- For any polyfunctional cell, suppose it is activated in makers \((i^*, j^*, k^*)\), with

\[ (i^*, j^*, k^*) \in \mathcal{I}^{(3)}_M = \{(i^*, j^*, k^*) \in \mathbb{N}^3 : 1 \leq i^* < j^* < k^* \leq 4\}. \]

For any \((i', j') \in \mathcal{S}\), either \(i' \in (i^*, j^*, k^*)\) or \(j' \in (i^*, j^*, k^*)\). Therefore, we expect the polyfunctional cell to belong to the polyfunctional cell-enriched set, denoted by

\[ C_{\text{poly}} = \cap_{(i,j) \in \mathcal{S}} C_{ij}. \]

12
For any bifunctional cell, suppose it is activated in markers \((i^*, j^*)\), with
\[
(i^*, j^*) \in \mathcal{T}^{(2)}_M = \{(i^*, j^*) : (i^*, j^*) \in S\} = S,
\]
We expect the bifunctional cell to belong to \(\cap_{(i,j) \in S^{(2)}_{(i^*, j^*)}} C_{ij}\) with
\[
S^{(2)}_{(i^*, j^*)} = S \setminus \{(i', j') : i' < j', i', j' \notin \{i^*, j^*\}\}.
\]  
(8)
Because \((i^*, j^*)\) can take any values in \(\mathcal{T}^{(2)}_M\), we denoted the bifunctional cell-enriched set by
\[
C_{bi} = \bigcup_{(i^*, j^*) \in S} \cap_{(i,j) \in S^{(2)}_{(i^*, j^*)}} C_{ij} \setminus C_{poly}.
\]
For any monofunctional cell, suppose it is activated in marker \(i^*\), with
\[
i^* \in \mathcal{T}^{(1)}_M = \{1, 2, 3, 4\}.
\]
We expect them to belong to \(\cap_{(i,j) \in S^{(1)}_{i^*}} C_{ij}\) with
\[
S^{(1)}_{i^*} = \{(i', j') : i' = i^* \text{ or } j' = i^*\}.
\]  
(9)
Because \(i^*\) can take any values in \(\mathcal{T}^{(1)}_M\), we denoted the monofunctional cell-enriched set by
\[
C_{mono} = \bigcup_{i^* \in \mathcal{T}^{(1)}_M} \{(i', j') \in S : i' = i^* \text{ or } j' = i^*\} \setminus \{C_{bi} \cup C_{poly}\}.
\]
We would like to point out that \(C_{mono}, C_{bi}, \text{ and } C_{poly}\) may still contain non-activated or less activated cells. However, by taking intersections among sets, we already filtered out many cells so that the remaining sets were highly enriched in the monofunctional, bifunctional, and polyfunctional cells.

4.3. Analysis using TEAM

In practice, \(\Omega_{ij}^+\) and \(C_{ij}\) are unknown; thus, we used TEAM to infer them and subsequently estimate the set of monofunctional, bifunctional, and polyfunctional cells. To accomplish this, we apply a two-step procedure.

In step 1, we applied TEAM in a series of six sub-analyses. In each subanalysis, we applied TEAM to a bivariate subspace to identify the regions where cohort 2 density is higher. Our analyses compared \(N_1 \approx 2.4\) million cohort 1 cells with \(N_2 \approx 2.2\) million cohort 2 cells. Cohort 2 contains the activated T cells that recognize CMV pp65 antigen. If a T cell is activated, it will be over-expressed in the effector molecules whose concentrations were measured in the data. Our goal is to identify the regions where the activated cells are enriched. Specifically, we used the sequential partitioning algorithm with 148 bins along each dimension (total \(148^2 = 21,904\) bins, with \(n \approx 210\) cells in each bin). We also set the algorithm to stop after three layers. The nominal FDR is 0.05. As in the simulations, the choice of \(n\) was set such that \(n \approx \{2(N_1 + N_2)\}^{1/3}\). We also tried other settings of layers,
number of bins, and bin size to check the robustness of TEAM. A sensitivity analysis to compare the results with other choices of bin numbers $m$ (and thus bin size $n$) is in Section S6 of the Supplementary Materials.

In step 2, we further narrowed down the activated cells in cohort 2. For each subanalysis in step 1, we denote all the cells falling into the identified regions by $\hat{C}_{ij}$, where the marker pair $(i, j) \in S$ defines the subspace. Using all the $\hat{C}_{ij}$ across the six sub-analyses, we estimated the monofunctional, bifunctional, and polyfunctional cell-enriched sets, $\hat{C}_{\text{poly}}$, $\hat{C}_{\text{bi}}$, and $\hat{C}_{\text{mono}}$, which were defined in Section 4.2. The monofunctional, bifunctional, and polyfunctional cell-enriched sets comprised 1.8%, 0.16%, and 0.23% of the cohort 2 cells, respectively. The nonfunctional cell set (containing all other cells) comprised 97.9% of the cohort 2 cells.

We plotted the 1D functional protein marker densities in Supplementary Figure S2. All functional markers are over-expressed in activated T cells, in agreement with our intuitions. We plotted the 2D densities of the lineage (Figure 4A) and maturational (Figure 4B) protein marker expressions to compare the distribution of the monofunctional, bifunctional, polyfunctional cell-enriched sets against the nonfunctional T cells (cohort 1). In Figure 4A, CD3 is a T cell co-receptor, and CD4 and CD8 are cell surface proteins that characterize T cells in helper and cytotoxic subtypes, respectively. The first row of Figure 4A shows that CD3 is increasingly down-regulated in more highly activated cells. This is consistent with the internalization and down-regulation of the TCR upon TCR-ligand activation (Valitutti et al., 1997).

The “dump channel” multiplexed markers CD14/CD19/vAmine that share the same fluorochrome are used to exclude monocytes, macrophages, B cells and dead cells. The second row of Figure 4A shows the enrichment of CD4+CD8+ double positive T cells among the polyfunctional cells compared to the nonfunctional cells. The CD4-CD8+ T cells are enriched among either monofunctional or polyfunctional cells, but not apparent among bifunctional cells. In Figure 4B, CD27, CD57, and CD45RO are cell-surface proteins that are differentially expressed in naive, effector and memory T cells. We excluded CD4-CD8-double negative T cells to highlight only the well-known T cells. The first row of Figure 4B shows that the activated T cells are mostly naive (N) and central memory (CM) T cells, with some polyfunctional cells having an effector memory (EM) phenotype. The second row of Figure 4 shows that the terminal effect (TE) and effector (E) T cells are more enriched in polyfunctional T cells, compared to the monofunctional and bifunctional T cells.

4.4. Analysis using other competing methods

As a comparison, we applied the scan-statistic method in Walther et al. (2010) and MRS in Soriano and Ma (2017) to the same data. Unfortunately, the scan-statistic method in Walther et al. (2010) failed to complete the analysis within one week on a machine with a 2.10GHz Intel Xeon Gold 6252 single-core processor, while TEAM only took 15 minutes on the same machine. Each run of MRS up to six (default) or seven layers takes about 5 hours. We also ran MRS using the maximum-allowable tree depth of 14, in order to match the same level of high-resolution as the first layer of TEAM ($\log_2(21,904) \approx 14$). This analysis did not complete within one week on the same machine.
Figure 4: 2D lineage (panel A) and maturational (panel B) marker densities of the cells in the estimated monofunctional, bifunctional, and polyfunctional cell-enriched sets (displayed as greyscale contour lines) against the nonfunctional cells (cohort 1; displayed as filled greyscale contours). In panel B, we excluded CD4-CD8- double-negative T cells to focus on the well-known T cell subsets, which include the five respective maturational subsets (N: naïve, CM: central memory, EM: effector memory, TE: terminal effector, E: effector).
We applied MRS to our dataset under two settings: 1) using only information in the four functional markers, TNF-α, IL-2, IFN-γ, and CD107, and 2) using the information on all 11 markers. We ran MRS up to layer six (default) and seven and controlled the FDR of the discovered bins (same as “windows” defined in Soriano and Ma (2017)) at the 0.05 level.

By using only the four functional markers, MRS did not identify any bins harboring significantly differential density, regardless of the tree depth. By using all eleven markers, we identified several significant bins. Figure S1 in the Supplementary Materials displays the maximum a posteriori (MAP) tree with depths of six and seven. Red nodes marked by red squares (henceforth called MRS-significant bins) correspond to the significant bins where cohort 2 cells are more enriched.

We also discovered that most bins identified by a six-layer MRS algorithm and by a seven-layer MRS algorithm are non-overlapping. The area-based Jaccard index (see Section S6 in the Supplementary Material for definition) for cohort 2 cells is zero across all MRS-significant bins, except for those corresponding to nodes B6 and C7, which have a similarity index of 0.97. The discrepancy in these results is likely due to the different generative priors under each tree depth. As a result, MRS results can be highly variable with different tree depths. This introduces difficulty in interpreting the results. In contrast, each layer of TEAM is fully nested so the region identified by the six-layer TEAM must be the subsets of those identified by the seven-layer TEAM. Computationally, this also saves time by making the analysis additive: to add an extra layer of analysis, start from the existing results and run only one additional layer.

5. Numerical Experiments

5.1. Experiment settings

To demonstrate the performance of TEAM, we designed five simulation settings that mimic the flow cytometry analysis. Because we simulated the data, the true monofunctional, bifunctional, and polyfunctional cells are known.

Settings SE1-SE3 simulated the monofunctional cells that will express one cytokine after being challenged with antigen. As shown in Section 4, the distribution of monofunctional cells is usually the most similar to the distribution of the nonfunctional cells, and these distributions are the most challenging to be differentiated. In setting SE4, we simulate the bifunctional cells that are activated in two protein markers and applied the bivariate TEAM analysis to this setting. In setting SE5, we mimicked the EQAPOL data and simulated two cell cohorts. Cohort 2 consists of nonactivated, monofunctional, bifunctional, and polyfunctional cells whose expression are defined across four protein markers.

The details of settings SE1-SE5 are listed below. Here \( Y \) represents the cellular protein marker expression. Under all settings, we simulate only a very small proportion of cohort 2 cells to be activated so that the cohort 2 cell density only deviates from the cohort 1 cell density a little to increase the difficulty level of the problems. This also mimics real flow cytometry studies, because only a small proportion of cells are expected to be activated (as illustrated in Section 4). Figure 5 illustrates the protein marker distributions of two cohorts of cells under settings SE1-SE4.
SE1. **One-dimensional local shift.** We randomly generated $N_1 = N_2 = 1,474,560$ random samples in each cohort. Assume

$$Y \sim \begin{cases} 
0.97\mathcal{N}(0.4, 0.04^2) + 0.03\mathcal{N}(0.88, 0.01^2) & \text{in cohort 1} \\
0.97\mathcal{N}(0.4, 0.04^2) + 0.03\mathcal{N}(0.89, 0.01^2) & \text{in cohort 2.}
\end{cases}$$

SE2. **One-dimensional local dispersion difference.** We randomly generated $N_1 = N_2 = 1,474,560$ random samples in each cohort. Assume

$$Y \sim \begin{cases} 
0.97\mathcal{N}(0.4, 0.04^2) + 0.03\mathcal{N}(0.8, 0.02^2) & \text{in cohort 1} \\
0.97\mathcal{N}(0.4, 0.04^2) + 0.03\mathcal{N}(0.8, 0.03^2) & \text{in cohort 2.}
\end{cases}$$

SE3. **One-dimensional local shift plus dispersion difference.** We randomly generated $N_1 = N_2 = 1,474,560$ random samples in each cohort. Assume

$$Y \sim \begin{cases} 
0.97\mathcal{N}(0.4, 0.04^2) + 0.03\mathcal{N}(0.8, 0.04^2) & \text{in cohort 1} \\
0.97\mathcal{N}(0.4, 0.04^2) + 0.03\mathcal{N}(0.82, 0.05^2) & \text{in cohort 2.}
\end{cases}$$

SE4. **Two-dimensional Gaussian mixture with extra cohort 2 samples in some subregions.** We generated $N_1 = N_2 = 500,000$ random samples for each cohort. For all cohort 1 samples and 495,000 cohort 2 samples, $Y \sim \sum_{j=1}^{5} p_j \mathcal{N}(\mu_j, \Sigma_j)$, where

$$(p_1, \ldots, p_5) = (0.03, 0.14, 0.17, 0.26, 0.40),$$

$$\mu_1 = (9, 9)^t, \mu_2 = (3.4, 2.9)^t, \mu_3 = (8.7, -5.8)^t, \mu_4 = (-0.4, 3.5)^t, \mu_5 = (-6, -6.5)^t,$$

$$\Sigma_1 = \begin{pmatrix} 2.1 & 0.6 \\ 0.6 & 0.9 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.71 & 0.14 \\ 0.14 & 2.12 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 2.08 & 0.87 \\ 0.87 & 1.39 \end{pmatrix}, \Sigma_4 = \begin{pmatrix} 2.8 & 0.6 \\ 0.6 & 1.2 \end{pmatrix}, \Sigma_5 = \begin{pmatrix} 1.34 & -0.45 \\ -0.45 & 3.13 \end{pmatrix}.$$ 

For cohort 2, we generated an additional 5,000 activated samples (1% of cohort 2), $Y$ follows the uniform distribution in the $\mathbb{R}^2$ subregion

$$\{(6, 8.5) \otimes (9, 11)\} \cup \{(8, 10) \otimes (7, 8)\} \cup \{(10, 11) \otimes (9, 12)\}.$$ 

SE5. **Four-dimensional markers with nonactivated, monofunctional, bifunctional, and polyfunctional cells in cohort 2.** We generated $N_1 = 500,000$ cells in cohort 1 and $N_2 = 500,000$ cells in cohort 2. Among $N_2$ cohort 2 cells, $N_{21} = 98.4\% N_2$ are nonactivated cells and $N_{22} = 1.6\% N_2$ cells are activated. Each cell $c$ has four markers, denoted by $Y = (Y_1, Y_2, Y_3, Y_4)^t$. Depending on the cell type (nonactivated, monofunctional, bifunctional, and polyfunctional), $Y$ were drawn from different distributions.

(a) **Nonactivated:**

- $N_1$ cells in cohort 1 and the $N_{21}$ cells in cohort 2: For $j \in \{1, 2, 3, 4\}$, $Y_j \sim \chi^2(3)$ (i.e., the central $\chi^2$ distribution with 3 degrees of freedom).

(b) **Monofunctional:**
• 0.2\%N_2 cells in cohort 2: For any \(j \in \{2, 3, 4\}\), \(Y_j \sim \chi^2(3)\). \(Y_1 \sim g\), where \(g\) has the exponential distribution with scale parameter 1 and location parameter 20.

\[ g(x) = \exp(x - 20), \quad x \geq 20. \]

• 0.2\%N_2 cells in cohort 2: For \(j \in \{1, 3, 4\}\), \(Y_j \sim \chi^2(3)\). \(Y_2 \sim g\).

• 0.2\%N_2 cells in cohort 2: For \(j \in \{1, 2, 4\}\), \(Y_j \sim \chi^2(3)\). \(Y_3 \sim g\).

• 0.2\%N_2 cells in cohort 2: For \(j \in \{1, 2, 3\}\), \(Y_j \sim \chi^2(3)\). \(Y_4 \sim g\).

(c) Bifunctional:

• 0.2\%N_2 cells in cohort 2: \(Y_1, Y_2 \sim g\), \(Y_3, Y_4 \sim \chi^2(3)\).

• 0.2\%N_2 cells in cohort 2: \(Y_3, Y_4 \sim g\), \(Y_1, Y_2 \sim \chi^2(3)\).

(d) Polyfunctional:

• 0.2\%N_2 cells in cohort 2: For \(j \in \{1, 2, 3\}\), \(Y_j \sim g\); \(Y_4 \sim \chi^2(3)\).

• 0.2\%N_2 cells in cohort 2: For \(j \in \{1, 2, 3, 4\}\), \(Y_j \sim g\).

Setting SE5 simulated 4 markers. We followed the strategy in Section 4.2 to consider 6 subspaces spanned by the two-dimensional markers: \((Y_1, Y_2)\), \((Y_1, Y_3)\), \((Y_1, Y_4)\), \((Y_2, Y_3)\), \((Y_3, Y_4)\). In each subspace, based on the definition of \(g(x)\), the alternative region is \([^20, +\infty] \cap \mathbb{R}^+ \cup \{\mathbb{R}^+ \cap [20, +\infty]\}\).

5.2. TEAM’s performance in the numerical experiments

Under settings SE1–SE4, we directly applied TEAM to the simulated dataset. Under Setting SE5, we applied TEAM to each sub-analysis with markers \((i, j)\) with \(i < j\), similar to what described in Section 4.3.

Under settings SE1–SE3, we applied the one-dimensional sequential partition algorithm to obtain \(m^{(1)} = 2^{14}\) bins. Each bin contains about \(n = 180\) pooled samples. The number \(n\) is set such that \(n \approx \{2(N_1 + N_2)\}^{1/3}\). Under settings SE4 and SE5, each bin contains about 120 samples.

Under all settings, if a bin overlaps with the alternative region (where cohort 2 density is higher than cohort 1 density), the bin is called an alternative bin.

The number of layers \(L\) was chosen so that \(1000 \leq m^{(L)} < 2000\). Under settings SE1–SE3, \(m^{(5)} \approx 2^{10} = 1024\). Thus, we ran TEAM up to five layers. We also tried to run TEAM up to layer 6, but many times TEAM did not reject any additional hypotheses on layer 6, suggesting that running five layers might be sufficient. Under setting SE4 and for each sub-analysis under setting SE5, we applied the two-dimensional sequential partition algorithm to obtain \(m^{(1)} = 90^2 = 8100\) bins with about 120 samples in each bin. Based on the same \(L\) selection criterion, we ran TEAM up to layer 4.

Under settings SE1–SE4, the experiment was repeated 1000 times. Under setting SE5, because each experiment has 6 sub-analyses, the experiment was repeated 500 times. TEAM is very computationally efficient. One repetition only takes less than 5 seconds on a standard laptop with a 2.6 GHz Intel Core i5 processor and 16Gb memory. The R package of TEAM
is available on R CRAN. The vignette is available at https://gitlab.oit.duke.edu/jichunxie/xie-lab-softwareTEAM.

Figure 6 and Figure 7 show how TEAM performs when stopping at layer \( L = 1, 2, 3, 4, 5 \). Specifically, it shows the average realized false discovery proportion, the average number of false negatives, the average total discoveries, and their 95% confidence intervals under the 5 settings. Under setting SE5, the average performance across 6 analyses was reported. TEAM successfully controls the FDR under the desired level after 5 layers for settings SE1 - SE3, and after 4 layers for settings SE4 and SE5. Meanwhile, the average false negatives substantially go down as TEAM proceeds to higher layers. For example, under setting SE1, at the desired FDR level 0.05, the single-layer method misses more than 80 alternative bins (out of about 245 total alternative bins); a five-layer TEAM only misses about 40 alternative bins. The power substantially increases as the layer goes up. Clearly, TEAM will make more true rejections on higher layers. For the purpose of checking robustness, We also tried to change the number of samples contained in each bin. Instead of using \( n = 180 \), we tried multiple settings with \( n \in [80, 250] \), and the results of TEAM were similar.

For setting SE5, we applied TEAM in 6 sub-analyses, following the same strategy described in Section 4.3 to obtain the monofunctional, bifunctional, and polyfunctional cell-enriched sets. The union of these sets is called the activated cell-enriched sets. We compared our estimated enriched sets with the true sets in terms of classification accuracy. The sensitivity and specificity of the enriched sets are displayed in Table 1 for varying nominal FDR and numbers of layers. It is clear that sensitivities of the enriched sets increase as the nominal FDR and number of layers increase, while the specificity remains high and close to 100%. The polyfunctional cell-enriched set \( \hat{C}_{\text{poly}} \) has the highest sensitivity, followed by the bifunctional cell-enriched set \( \hat{C}_{\text{bi}} \), and finally, the monofunctional cell-enriched set \( \hat{C}_{\text{mono}} \). This is not surprising because polyfunctional cells are activated in three or four markers and thus show much stronger signals. Because we run TEAM across 6 bivariate sub-analyses, those regions likely to be enriched for polyfunctional cells have signals along both markers, and thus the differential density regions are more likely to be identified.

Figure 5: Density functions under SE1 - SE4. In the first three panels, the solid curve represents the PDF of cohort 1 (control) and the dashed curve represents PDF of cohort 2 (stimulated). In the fourth panel, the Gaussian mixture PDF of cohort 1 is illustrated by the heat-map, and the dashed rectangle illustrates the uniform distribution regions of the 1% stimulated samples in cohort 2.
Figure 6: Performance of TEAM under SE1 - SE4. The first row shows the average realized false discovery proportion with TEAM stopping at different layers ($L = 1, 2, 3, 4, 5$). The second row shows the average number of false negatives when TEAM stops at different layers. As a reference, the average numbers of alternative leaves under SE1 - SE4 are 245, 160, 329, 130. The third row shows the average number of total discoveries of TEAM when stopping at different layers. For each performance measure, we present the 95% confidence intervals, calculated as the 2.5% and 97.5% quantiles over the 1000 repetitions.

Figure 7: Performance of TEAM under SE5. From left to right, the panels show the average realized false discovery proportion, average number of false negatives, and average number of rejections when TEAM stops at different layers ($L = 1, 2, 3, 4$) across 6 sub-analyses and 500 repetitions. As a reference, the average numbers of alternative leaves under SE5 is 179. For each performance measure, we present the 95% confidence intervals, calculated as the 2.5% and 97.5% quantiles averaged over the 500 repetitions of 6 sub-analyses.
Table 1: Classification accuracy for cohort 2 activation classes across varying nominal FDR and layer numbers for simulation setting S5. Performance is displayed in the format of (sensitivity, specificity), omitting the percentage marker %.

<table>
<thead>
<tr>
<th>Nominal FDR</th>
<th>Layer</th>
<th>Cohort 2 Cell Type</th>
<th>Activated</th>
<th>Monofunctional</th>
<th>Bifunctional</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>1</td>
<td>(30.4, 100.0)</td>
<td>(0.3, 99.8)</td>
<td>(0.3, 99.9)</td>
<td>(51.8, 100.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>(40.5, 99.9)</td>
<td>(2.4, 99.7)</td>
<td>(4.7, 99.9)</td>
<td>(66.3, 100.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>(47.1, 99.8)</td>
<td>(8.8, 99.5)</td>
<td>(13.3, 99.9)</td>
<td>(85.5, 100.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>(51.1, 99.6)</td>
<td>(15.8, 99.3)</td>
<td>(15.5, 100.0)</td>
<td>(88.4, 100.0)</td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>1</td>
<td>(33.3, 100.0)</td>
<td>(0.7, 99.7)</td>
<td>(0.8, 99.9)</td>
<td>(53.6, 100.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>(42.9, 99.9)</td>
<td>(3.6, 99.7)</td>
<td>(7.1, 99.9)</td>
<td>(70.3, 100.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>(48.6, 99.7)</td>
<td>(10.1, 99.5)</td>
<td>(14.8, 99.9)</td>
<td>(85.9, 100.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>(52.7, 99.5)</td>
<td>(17.1, 99.3)</td>
<td>(16.8, 100.0)</td>
<td>(88.6, 100.0)</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>1</td>
<td>(35.6, 100.0)</td>
<td>(1.1, 99.7)</td>
<td>(1.5, 99.9)</td>
<td>(55.4, 100.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>(44.9, 99.9)</td>
<td>(4.9, 99.6)</td>
<td>(9.4, 99.9)</td>
<td>(73.3, 100.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>(49.8, 99.7)</td>
<td>(11.1, 99.5)</td>
<td>(16.0, 99.9)</td>
<td>(85.9, 100.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>(53.9, 99.5)</td>
<td>(18.3, 99.2)</td>
<td>(18.3, 100.0)</td>
<td>(88.7, 100.0)</td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>1</td>
<td>(37.5, 100.0)</td>
<td>(1.5, 99.7)</td>
<td>(2.3, 99.9)</td>
<td>(57.5, 100.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>(46.3, 99.8)</td>
<td>(6.1, 99.6)</td>
<td>(11.3, 99.9)</td>
<td>(75.2, 100.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>(50.9, 99.7)</td>
<td>(12.2, 99.4)</td>
<td>(17.2, 99.9)</td>
<td>(86.5, 100.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>(55.0, 99.5)</td>
<td>(19.6, 99.2)</td>
<td>(19.7, 100.0)</td>
<td>(89.0, 100.0)</td>
<td></td>
</tr>
</tbody>
</table>
5.3. Comparison with multi-resolution scanning

We compared the performance of TEAM with the MRS method proposed by Soriano and Ma (2017). When we ran MRS, we set the MRS-type FDR at the levels 0.05, 0.10, 0.15, and 0.20. It is important to clarify that this MRS-type FDR is not the same as the FDR defined on those hypotheses on the finest resolution level. See explanations in Section 1. Because MRS is a two-sided testing method embedded in the partition tree, we perform the following procedure to obtain comparable effective discoveries and conduct a fair comparison. First, we ran MRS up to 14 layers under settings SE1-SE3 and 13 layers under setting SE4. Second, we summarized discoveries made in the last 5 layers under settings SE1-SE3 and the last 4 layers under setting SE4. This was to make sure that the layers we compared in TEAM and MRS had comparable resolution levels. Third, among those discoveries, we summarized those with effect size favoring cohort 2. This is because MRS performs a two-sided test while TEAM performs a one-sided test. Fourth, we mapped those discoveries to the finest resolution layer (the last layer). We call those discovered bins the effective discoveries. This is to mimic the mapping step in TEAM to calculate the realized FDP on the finest layer.

In Table 2, we report the percentage of repetitions where MRS made at least one rejection, and among those repetitions with one or more rejections, the average number of rejections and the realized FDP when mapping discoveries to the finest resolution layer. This realized FDP has the same definition as the TEAM algorithm. Clearly at fine-resolution levels, the performance of MRS is not satisfying. Under settings SE1-SE3, MRS fails to identify any bins in the last 5 fine-resolution layers in most repetitions. This is because we simulated very challenging settings. However, these settings are common in flow cytometry analysis. Among those repetitions with identified bins, the realized FDP is very high. Under settings SE4 and SE5, MRS can identify bins in the last 4 fine-resolution layers in most repetitions, but the corresponding realized FDP is extremely high. The fundamental reason for MRS’s poor performance under these settings is because MRS is a tool designed for visualizing global multi-dimensional distribution differentiation and automatically searching for splitting direction; it is not designed for pinpointing fine-resolution differential density regions. Thus, while MRS has its own advantages and will be appropriate for certain application scenarios, it is not suitable for pinpointing the small density differential regions that harbor the activated cells.

6. Discussion

In this paper, we proposed a novel method TEAM to pinpoint the protein marker density differential regions based on flow cytometry data. With additional filtering steps, TEAM can be used to identify the activated T cells enriched sets. In this paper, we applied TEAM to a flow cytometry study to identify the monofunctional, bifunctional, and polyfunctional T cell-enriched sets. We also characterized the activated T cells, while other competing methods failed to provide such comprehensive information.

We recommend applying TEAM when the protein marker is known to be differentially expressed or “functional”. If a protein marker is actually not differentially expressed under two conditions, TEAM may lead to an inflated FDR. Please see Condition C3 in Section S2.
Table 2: Performance of MRS in SE1-SE5. Each setting has 1000 repetitions. “% Rej” is percentage of repetitions where MRS made at least one effective discovery; “Avg. Rej” is, among those repetitions with one or more effective discoveries, the average number of discovered bins when MRS’s effective discoveries are mapped to the finest resolution layer; “Avg. FDP” is, among those repetitions with one or more effective discoveries, the average realized false discovery proportion after MRS’s effective rejection are mapped to the finest resolution layer.

<table>
<thead>
<tr>
<th></th>
<th>% Rej</th>
<th>Avg Rej</th>
<th>Avg FDP</th>
<th>% Rej</th>
<th>Avg Rej</th>
<th>Avg FDP</th>
<th>% Rej</th>
<th>Avg Rej</th>
<th>Avg FDP</th>
<th>% Rej</th>
<th>Avg Rej</th>
<th>Avg FDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE1</td>
<td>0.8</td>
<td>10</td>
<td>0.52</td>
<td>3.0</td>
<td>8.5</td>
<td>0.44</td>
<td>9.8</td>
<td>980.5</td>
<td>0.54</td>
<td>25.6</td>
<td>3999.9</td>
<td>0.63</td>
</tr>
<tr>
<td>SE2</td>
<td>0.9</td>
<td>1.1</td>
<td>0.11</td>
<td>1.4</td>
<td>1.3</td>
<td>0.14</td>
<td>2.7</td>
<td>1.5</td>
<td>0.16</td>
<td>5.4</td>
<td>1.4</td>
<td>0.22</td>
</tr>
<tr>
<td>SE3</td>
<td>1.9</td>
<td>1.3</td>
<td>0.16</td>
<td>3.3</td>
<td>1.4</td>
<td>0.17</td>
<td>7.6</td>
<td>629</td>
<td>0.23</td>
<td>16.4</td>
<td>2147.2</td>
<td>0.31</td>
</tr>
<tr>
<td>SE4</td>
<td>95.8</td>
<td>297.7</td>
<td>0.69</td>
<td>96.9</td>
<td>285.8</td>
<td>0.68</td>
<td>97.3</td>
<td>252.7</td>
<td>0.68</td>
<td>98.2</td>
<td>238.9</td>
<td>0.67</td>
</tr>
<tr>
<td>SE5</td>
<td>86.2</td>
<td>1871.9</td>
<td>0.83</td>
<td>94.2</td>
<td>1142.2</td>
<td>0.71</td>
<td>97.7</td>
<td>665.7</td>
<td>0.55</td>
<td>98.9</td>
<td>394.8</td>
<td>0.41</td>
</tr>
</tbody>
</table>

in the supplementary materials. If the differentially expressed markers are unknown, other single-cell differential expression tools can be applied. See Wang et al. (2019) with a review and comparison of 11 such tools.

Under the complicated scenarios, even after preprocessing and normalization steps, the maker expression density functions from the two cohorts may still not be well aligned. Hahne et al. (2010a) have provided some examples where the two densities are not well-aligned in the negative peak regions. Thus, TEAM may identify the negative peak regions where the activated cells are less likely to reside. However, we can use domain knowledge to screen the regions and improve the accuracy.

Although we focused on applying TEAM to flow cytometry analysis in this paper, TEAM can be used to draw insights from data in other contexts. For example, TEAM can be applied to single-cell RNA sequencing (scRNA-seq) data to identify the activated cells after stimulation. We can use domain knowledge or differential gene expression analysis to select the signature markers and then apply TEAM in multiple sub-analyses to filter and characterize the activated cells. TEAM can also be applied to CHIP-seq or ATAC-seq data to compare the read counts mapping to the genome and detect the regions with differential epigenomic signals (such as differential open chromatin regions, differential footprint, et al.) We will explore the application of TEAM on other application contexts in the future.

Acknowledgements

Flow cytometric data generation was supported in whole or part through an EQAPOL collaboration with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Contract Number HHSN272201700061C.

John Pura’s research was supported by NSF DGE 1545220 and the NIH training grant T32HL079896. Cliburn Chan’s research was supported by EQAPOL Contract Number
HHSN272201700061C, and the Duke University Center for AIDS Research (CFAR), and an NIH-funded program (5P30 AI064518). Jichun Xie’s research was sponsored by her startup fund from Duke University.

We thank the Editors, the Associated Editors, and the reviewers for providing valuable comments that helped us substantially improve the paper’s quality.
References


