

Department of Electrical Engineering
Columbia University

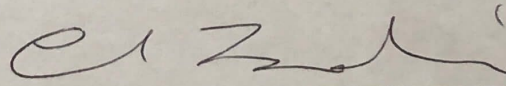
Curricular Practical Training Report
Coversheet

Name: ADITYA SINHA

Company Name: MEMORIAL SLOAN KETTERING CANCER CENTER

Supervisor Name: DR. CHRISTINA LESLIE

Supervisor's Email: cleslie@cbio.mskcc.org

Supervisor's Signature: 

Automating Flow Cytometry Analysis

Aditya Sinha

August 18, 2019

1 Introduction

1.1 Background & Company Description

With the advent of the modern era of healthcare, infectious diseases have been all but eradicated in most developed countries. With that milestone behind us, the next challenge is to tackle genetic diseases, ie. those caused by an unwanted mutation in normal gene expression. Within these, cancer is especially malicious, with over 1.5 million new cases and over half a million deaths each year in the United States alone. The lifetime probability of any individual being diagnosed at some point in their lives is about 40%, which is one of the major motivations behind research in cancer biology.

Here at the Sloan Kettering Institute (SKI) at Memorial Sloan Kettering Cancer Center, we seek to understand the biological and molecular mechanisms of various cancers, as well as their variability among patients. Our efforts and discoveries drive clinical progress in fields like immunotherapy and drug discovery at MSK and beyond. The wide spectrum of research going on at SKI makes it hard to provide an encompassing blanket statement, but I've been working in the field of computational biology at the Christina Leslie Lab this summer, and our lab develops computational methods to study biological systems from a global and data-driven perspective. We seek to exploit the diverse genomic data collected from various methods to understand molecular networks underlying fundamental cellular processes. Our algorithmic methods draw on machine learning as a backbone to support analysis of high dimensional genomic data.

My job description and assigned project involved developing machine learning methodologies for analysis of single cell flow cytometry data in tumor and immune cells of patients suffering from Renal Cell Carcinoma. This is relevant to my course program at Columbia University as my major is Electrical Engineering, with a concentration in Systems Biology & Neuroengineering. My work at SKI seeks to use the algorithms and computational skills developed as an electrical engineer in the field of systems biology - moving towards my life goal of trying to understand the mechanisms of how we work better.

2 Analysis of Flow Cytometry Data

Flow cytometry (Fig. 1) is a technique used to detect and measure physical and chemical characteristics of a population of cells or particles. A sample containing cells or particles is

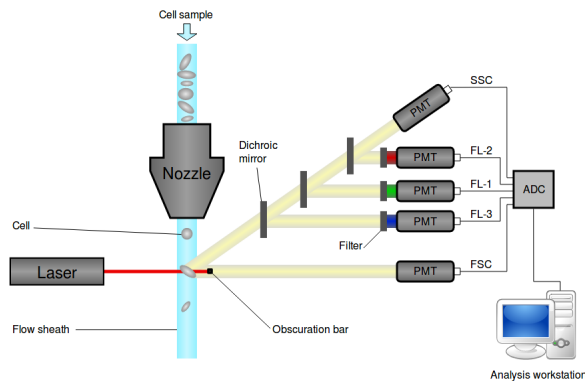


Figure 1: Structure and working of a flow cytometer

suspended in a fluid and injected into the flow cytometer instrument. The sample is focused to flow one cell at a time through a laser beam and the light scattered is characteristic to the cells and their components. Cells are first labeled with fluorescent markers using an antibody staining panel, which allows us to record fluorescence values of different protein markers, along with the basic Forward and Side Scatter (FSC - measure of cell size, SSC - cell complexity).

2.1 Classical Flow Analysis

The pipeline for classical flow analysis (Fig. 2) constitutes of consecutive application of manual gates on the protein markers to filter out the flow data and find populations of cells in the samples (T cells, NK cells, ILCs, myeloid populations for example). By gating, we mean applying filtering thresholds to 2D slices of the 13 colour flow data cube in a successive manner. Think of this as taking a chisel and digging out a section of the multidimensional hypercube by considering two dimensions at a time.

Preliminary cleaning of data involves using the compensated data (the instrument compensates for bleed through) gating for live, CD45+ singlets.

- Live - filters out dead cells
- CD45+ - marker for all immune cells
- Singlets - the instrument ideally flows one cell at a time, but sometimes cells stick together and we need to filter those readings out

The gating strategy is inspired from what we expect biologically and it allows us to see populations at various granularities. The disadvantage of this technique, however, is that this sort of stringent viewing can tunnel vision us and we might miss out on some populations. This is what motivates us to look for a more global approach at dividing the sample into various known populations.

2.2 Proposed Analysis Pipeline

For the purpose of comparison of population statistics, the proposed pipeline has been applied to the same sample for which we have manual gating done - tumor sample of a chromophobe RCC (chRCC is a variant) patient, stained with a lineage panel (tumor lineage is a common assay to determine the development and differentiation of various cell populations - refer to Fig. 3). The proposed pipeline for analysis is as follows:

- Use the compensated data with preliminary filtering for live, CD45+ singlets done
- Perform a logicle transform on the data. Since the data is of fluorescences, it is in orders of magnitude and it needs to be scaled for it to be meaningful. The logicle transform is like log, except it also deals with negative values, mapping them close to zero.
- Perform k-Nearest Neighbours (kNN) and then louvain clustering on the nearest neighbour sets.
- Visualise the 13 dimensional data using t-SNE embedding
- Vary the kNN and the resolution parameters till you get meaningful clusters (populations)
- The gene expression of each cluster can be analyzed by constructing violin plots. This enables classification and naming of the populations based on marker activity.
- Vary the resolution to see populations and subpopulations at various granularities. Low resolution: less clusters, broad populations; High resolution: more clusters, subpopulations.
- Record cell counts and compare them with classical flow analysis to validate. Furthermore, can find new uncharacterized populations from the clustering.

3 Results

So as to not get overwhelmed by plots, let us look at one particular resolution and its corresponding violin plots for the 11 markers in the tumor sample of chRCC-638 (out of 13, CD45 and Live/Dead have already been used for preliminary gating). In Fig. 4, we see 7 clusters in the t-SNE. The marker expressions for these clusters can be seen as violin plots in Fig. 5, providing us with insight into what these individual populations are.

An example characterization of this would look somewhat like this: clusters 1,3,5,7 are TCRab positive and are hence the T-cells, clusters 1 and 7 are CD8 T-cells (killer T), with a differentiation based on GzmB expression (cytotoxicity), whereas clusters 3 and 5 are CD4 T cells (helper T), again in two states based on their GzmB expression. Cluster 4 is CD56dim CD16+ GzmB+, with low expression of CD4, CD8 and TCRab and hence comprises of Natural Killer (NK) cells, whereas cluster 6 is CD56bright and tissue resident (CD49a+,CD103+) and thus consists of Innate Lymphocytes (ILCs). Cluster 2 is what we

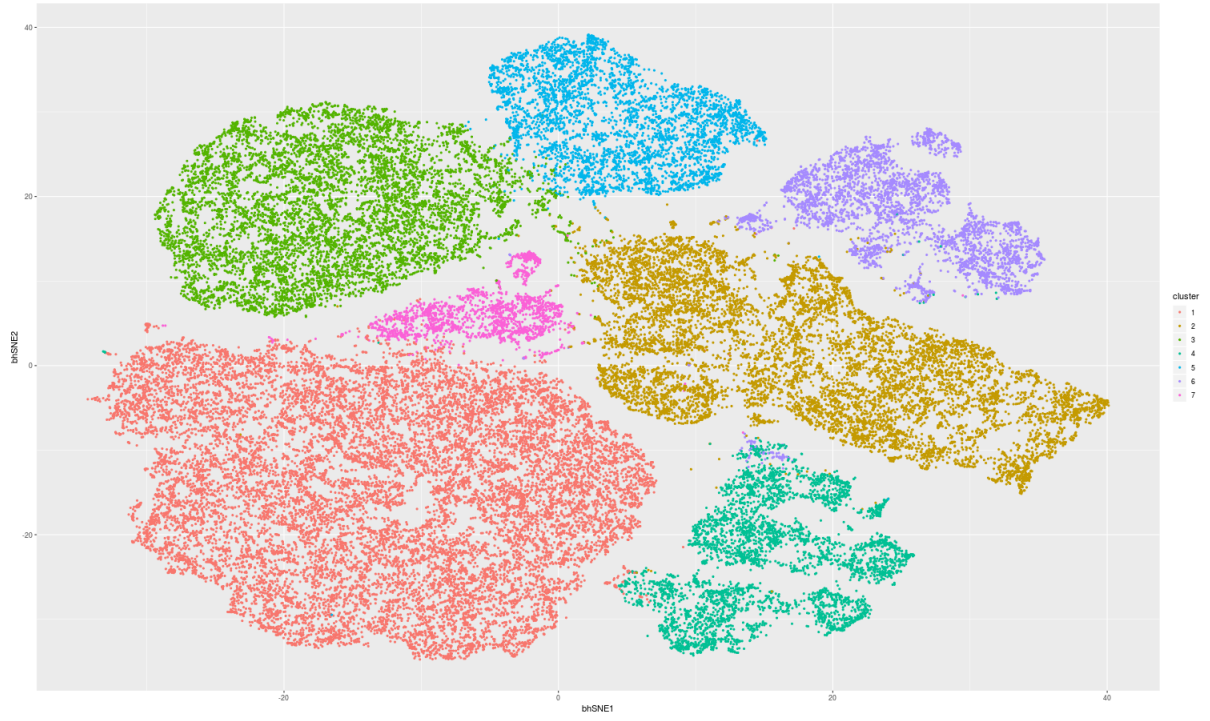


Figure 4: Louvain clustering/t-SNE with 7 clusters

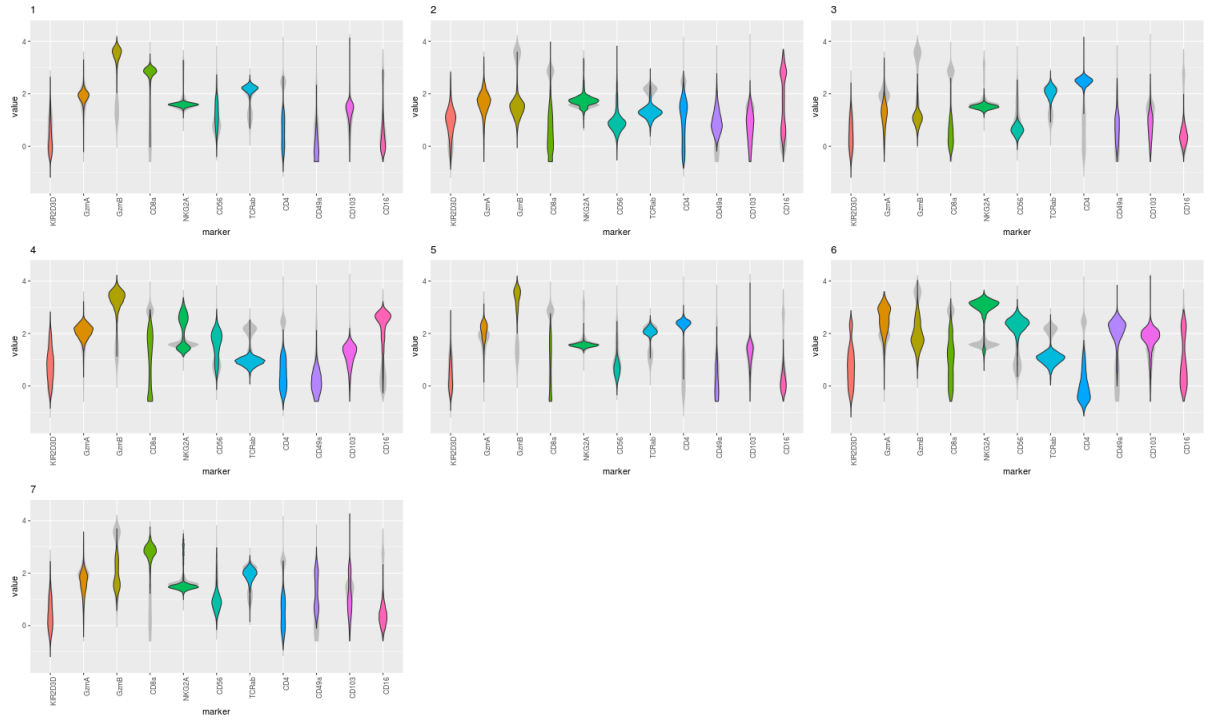


Figure 5: Violin plots for protein marker expression of each cluster

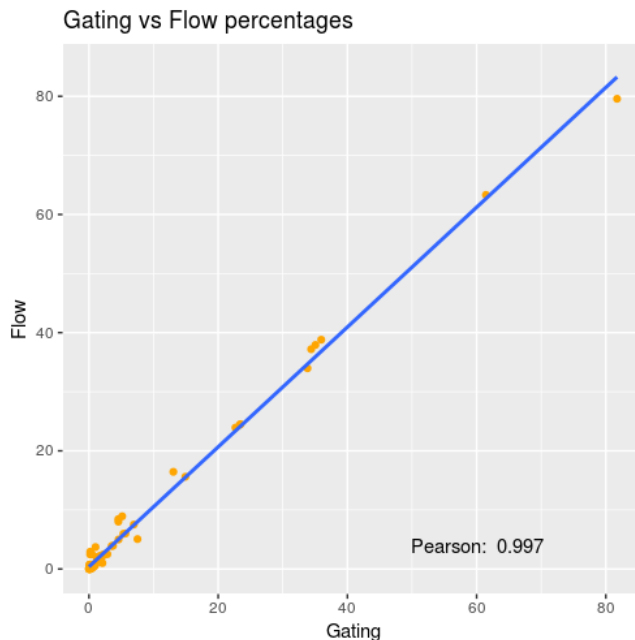


Figure 6: Correlation between manual gating and clustering percentages

affectionately call the "confusing cluster", containing of populations we can't classify, mostly because this panel does not have markers for them (myeloid populations, for example).

The cell counts for these various populations and subpopulations at various granularities were recorded as a total percentage of CD45+ cells and were then compared with those from manual gating analysis. In Fig. 6, we can see that the recorded percentages are highly correlated, with a Pearson correlation of 0.997.

4 Discussion & Comparison

In essence, the difference between classical flow analysis and the new proposed method is that in the former, you have a 13-dimensional brick from which you methodically chisel out your favourite population, whereas in the latter, you take that brick and smash it on the floor, hoping to see everything at once. There are disadvantages to both of these. With manual gating, we can have a more precise look at marker activities and divide the populations based on biological plausibility, but that can tunnel-vision us and make us miss out on populations. Moreover, the gating strategy has to be decided for each panel and the thresholds for gating are arbitrary, based on what the immunologist feels is high/low. With the proposed clustering+tSNE method, we can visualize all populations at once, cross-reference them to the ones noted, and also find new populations that might have been missed out. Since this method does not rely on arbitrary thresholds, but on multimodal proximity to cluster populations, this might give us more sensible divisions. However, this technique may be imprecise in quantifying the populations and sometimes the clusters formed may not have biological meaning, creating bimodal artifacts out of markers that are supposed to be unimodal. An optimal pipeline likely exists in the combination of the two.

4.1 Future Work

The future work on this flow project would explore the following avenues:

- Generalizing the analysis to the entire cohort of patients and validating results
- Dealing with batch effect among patients
- Combination of clustering and gating to obtain a biologically more accurate characterization
- Looking for subpopulations that have been previously uncharacterized and showing their robustness
- Giving the data of these populations back to the immunologists for validation by further experiments

If we are successful in our efforts, it would be really cool, as it would allow us to rescue all this flow data. It would not only support the current scRNA-Seq analysis, but might prove to be a novel method for a more end-to-end, more global, more panoramic view of flow data.